

Proceedings of the Conference on Research Data Infrastructure

1 | 2023

1st Conference on Research Data Infrastructure (CoRDI)

Connecting Communities

Karlsruhe (Germany), 12 - 14 September 2023

Editors

York Sure-Vetter

Carole Goble



Proceedings of the Conference on Research Data Infrastructure

Research data form the basis for knowledge and innovation throughout all scientific disciplines. They play a fundamental role in the progress of our society. The key to using these data treasures is an effective infrastructure. With the first edition of the Conference on Research Data Infrastructure from 12 to 14 September 2023, the Association German National Research Data Infrastructure (NFDI) is initiating a conference that will focus on establishing interdisciplinary research data management (RDM).

ISSN (online): 2941-296X

TIB
OPEN
PUBLISHING

The The Proceedings of the Conference on Research Data Infrastructure are published by TIB Open Publishing (Technische Informationsbibliothek) on behalf of Nationale Forschungsdateninfrastruktur (NFDI) e.V.



All contributions are distributed under the Creative Commons Attribution 4.0 International License.

About the Conference on Research Data Infrastructure

Research data form the basis for knowledge and innovation throughout all scientific disciplines. They play a fundamental role in the progress of our society. The key to using these data treasures is an effective infrastructure. With the first edition of the Conference on Research Data Infrastructure from 12 to 14 September 2023, the Association German National Research Data Infrastructure (NFDI) is initiating a conference that focuses on establishing interdisciplinary research data management (RDM).

Under the theme Connecting Communities, national and international stakeholders from all research fields as well as from the infrastructure sector are invited to present their contributions to an excellent RDM of the future and to exchange information about the latest developments. NFDI is organizing the conference in cooperation with the Karlsruhe Institute of Technology (KIT). Over the course of three days, topics related to RDM and the joint development of an effective research data infrastructure for Germany and beyond are examined from a wide variety of perspectives. The Conference on Research Data Infrastructure (CoRDI) stands for more comprehensive knowledge through better use of research data, for innovations and the resulting social benefits.

Call for Papers and Call for Posters

It was possible to submit extended papers as part of a **Call for Papers**. Accepted submissions from this call are presented as talks or posters at the conference. They are part of the Proceedings of the Conference on Research Data Infrastructure.

There was also a **Call for Posters**. The posters that were accepted in this call can be published and found on Zenodo: <https://zenodo.org/communities/cordi-2023/>

Conference Tracks

Four disciplinary tracks address infrastructure aspects relevant for the individual disciplines:

- **Humanities and Social Sciences**
- **Natural Sciences**
- **Life Sciences**
- **Engineering Sciences**

Six cross-disciplinary tracks address more specific topics regarding the delivery of RDM infrastructure:

- **Enabling RDM (incl. software):** The aim of the contributions in this track is to develop infrastructure and software components that can be used by the entire community and ensure interoperability so that network effects can be systematically exploited. This includes infrastructure for data analysis such as workflow platforms.
- **Harmonising RDM:** This track is about (meta-)data, terminologies and provenance and how research data management can be harmonised on a broad basis across organisations and disciplines.
- **Securing RDM:** This track addresses ethical, legal and social issues associated with collecting and working with research data.
- **Spreading RDM:** The contributions to the track all deal with the multi-faceted topic of data literacy which plays a key role in solving current and future societal challenges and in understanding digital culture.

- **Linking RDM:** This track focuses on how research data can be shared as freely and openly as possible between the domains of science, business and society, taking into account the interests of different stakeholders.
- **Connecting RDM:** This track has a focus on how international and national Research Data Infrastructures can be synergistically connected with each other.

Keynotes

Christine Borgman

(University of California, Los Angeles): Knowledge Infrastructures: The Invisible Foundation of Research Data Or, How Infrastructure Connects and Disconnects Research Communities

Tuesday, 12 September, 11.30 – 12.30 h

Abstract

Implicit in investments in research data infrastructure is the assumption that data are valuable entities worth preserving, stewarding, sharing, and reusing. This value proposition also implies that research data are useful to others and that others will reuse those data. However, neither outcome is assured. Data practices are local, varying from field to field, individual to individual, and country to country. As the number and variety of research partners expands, so do the difficulties of sharing, reusing, and sustaining access to data. Efforts to develop global research infrastructures are hindered by communities' lack of agreement on data management practices –or on what constitutes 'research data.' This talk argues for a broader focus on knowledge infrastructures, which are robust networks of people, artifacts, and institutions for producing, exchanging, and sustaining knowledge. Technical aspects of infrastructure, from persistent identifiers to compute capacity and storage, are easier to address than are social aspects, such as data stewardship, trust, governance, economics, infrastructure, standards, and science policy. Infrastructures can connect communities when they support local practices, and disconnect communities when they create incompatible silos. Examples are drawn from several decades of empirical research with research communities in environmental sciences, sensor networks, astronomy, biomedicine, social sciences, and digital humanities.

Speaker

Christine L. Borgman is Distinguished Research Professor in Information Studies at the University of California, Los Angeles. She is recognized internationally for her research in information and computer science, data science, communication, digital humanities, privacy, and law. Her current research focuses on knowledge infrastructures, scientific data practices, and open science. Among her publications are three award-winning monographs from MIT Press: *Big Data, Little Data, No Data: Scholarship in the Networked World* (2015); *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (2007); and *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World* (2000). She has held visiting scholar posts at Oxford, Harvard, Lund, Budapest Economic Sciences, Eotvos Lorand, and Loughborough universities, and DANS (Netherlands). Professor Borgman is a Fellow of the [American Association for the Advancement of Science](#) and of the [Association for Computing Machinery](#).

Julia Janssen

(Artist and researcher): Public evening lecture: Behind the Click

Tuesday, 12 September, 19.30 – 20.30 h

Abstract

Julia Janssen is an artist who researches the influence of digitalisation on our society. She makes the challenges we face with data, AI and technology tangible in interactive and performative installations. She covers topics like data profiling, bias in algorithms, informed

consent and digital civic rights. How do we deal with fairness, equality, autonomy, freedom and democracy in a data-driven society? In this talk, Janssen takes you on a journey behind the surface of the internet; A visual presentation on her research and ideas to discover what happens behind the click.

Speaker

Julia Janssen is an artist, designer, researcher and speaker. In her work she creates awareness about the impact of technology and digitization on society. Sinds March 2022 she is ambassador for [Stichting Data Bescherming Nederland](#).

In 2016 she graduated from the ArtEZ School of the Arts in Graphic Design. During her studies, she became interested in data and digitization and worked on several projects that highlight our relationship with technology. With her graduation work, she won the [Crypto Design Award](#) and she has since devoted her art to data sovereignty.

Julia translates scientific insights into accessible design giving her audience a peek behind the internet's surface. By making the complexity of information technology understandable, she builds a movement that strives for data sovereignty. Ultimately, everyone must consciously choose who knows what about him or her on the internet. This philosophy is reflected in several projects: she develops and designs projects about the lack of ownership, control, and transparency over personal data, the capital in online behavioral patterns, the changing definition of privacy, and the future of digital identity.

Mark D Wilkinson

(Isaac Peral Senior Researcher, Centre for Plant Biotechnology and Genomics, Universidad Politécnica de Madrid): A Series of FAIR Vignettes

Thursday, 14 September, 10 – 11 h

Abstract

The activities required to achieve “FAIRness” span a wide range of very distinct expert domains, including library sciences, data and knowledge representation, semantics, Web communication, software development, standards and protocols, licensing, ethics/privacy/consent, and agent-based negotiation. In this presentation, Mark D Wilkinson will try to appeal to the various communities in the CoRDI audience by telling a series of stories that focus on different pieces of the larger FAIR puzzle. These will include some thoughts – and second-thoughts! – about the Principles themselves, as well as observations of the benefits of FAIR in-practice. Conversely, he will also provide examples of FAIRness challenges which continue to evade robust solutions despite the best efforts of FAIR practitioners, and drill-down into the technologies and/or behaviors that are creating these barriers.

Speaker

Mark D Wilkinson has a B.Sc.(Hons) in Genetics from the University of Alberta, and a Ph.D. in Botany from the University of British Columbia. He spent four years at the Max Planck Institut für Züchtungsforschung in Köln, Germany, pursuing studies in a mix of plant molecular and developmental biology and bioinformatics. He then did a research associateship at the Plant Biotechnology Institute of the National Research Council Canada, focusing on the problem of biological data representation and integration for the purposes of automated data mining. In the subsequent 20+ years, his laboratory has focused on designing biomedical data/tool representation, discovery, and automated reuse infrastructures – what are now called “FAIR Data” infrastructures. He is the lead author of the primary FAIR Data Principles paper, and lead author on the first paper describing a complete implementation of those principles over legacy data. He is a founding member of the FAIR Metrics working group, tasked with defining the precise, measurable behaviors that FAIR resources should exhibit, and the author of the first software application capable of a fully-automated and objective evaluation of “FAIRness”. He is co-Chair of the EOSC Task Force on FAIR Metrics and Data Quality, and is founder of a

spin-off company, FAIR Data Systems S.L., that provides consulting, training, and customized software solutions that help clients become FAIR.

Programme

Tuesday, 12.09.2023

10:00-11:00h	Registration Audimax Foyer (Building 30.95)			
11:00-11:30h	Welcome & Opening: Mario Brandenburg (Parlamentarischer Staatssekretär bei der Bundesministerin für Bildung und Forschung) Kerstin Schill (DFG-Vizepräsidentin) Ute Gunsenheimer (Secretary General, EOSC) Kora Kristof (Vizepräsidentin Digitalisierung und Nachhaltigkeit, KIT) Carole Goble (CoRDI Programme Chair, Univ. of Manchester & ELIXIR-UK), York Sure-Vetter (CoRDI General Chair, NFDI-Direktor & KIT) Audimax (Building 30.95)			
11:30-12:30h	Keynote: Christine Borgman , University of California, Los Angeles “Knowledge Infrastructures: The Invisible Foundation of Research Data Or, How Infrastructure Connects and Disconnects Research Communities” More Information Audimax (Building 30.95)			
12:30-14:00h	Lunch Audimax Foyer (Building 30.95)			
14:00-15:30h	Humanities & Social Sciences I Audimax (Building 30.95)	Natural Sciences I Großer Hörsaal (Building 10.50)	Life Sciences Kleiner Hörsaal (Building 10.50)	DFG Pecha Café* Festsaal (Adenauerring 7)
15:30-16:00h	Coffee break Audimax Foyer (Building 30.95)			
16:00-17:30h	Humanities & Social Sciences II Audimax (Building 30.95)	Natural Sciences II Großer Hörsaal (Building 10.50)	Engineering Sciences Kleiner Hörsaal (Building 10.50)	DFG Pecha Café* Festsaal (Adenauerring 7)

17:30-19:00h	Poster session I (Posters via Call for Posters) More information Audimax Foyer (Building 30.95)
19:00-19:30h	Break
19:30-20:30h	Public Keynote: Julia Janssen , artist and researcher “Behind the Click” More Information Audimax (Building 30.95)

* This is a continuous event. Participation in the parts before and after the break is optimal. The event will be held in German.

Session Details

Early afternoon (14:00 – 15:30h)

Humanities & Social Sciences I (Audimax (Building 30.95))

- P. Kamocki; E. Hinrichs; S. Springer; P. Leinen; A. Witt; D. Zechmann
Open Science and Language Data: Expectations vs. Reality: The Role of Research Data Infrastructures
- F. Thiery; A. Mees; B. Weisser; F. Schäfer; S. Baars; S. Nolte; H. Senst; P. von Rummel
Object-related Research Data Workflows within NFDI4Objects and beyond
- M. Fichtner; R. Nasarek; T. Wiesing
WissKI: A Virtual Research Environment based on Drupal
- S. Lieber; A. Van Camp; D. De Witte; E. Coudyzer; E. Buelinckx; E. Angenon; H. Lowagie; J. Birkholz; K. Lasaracina
MetaBelgica Project: A Linked Data Infrastructure Between Federal Scientific Institutes in Belgium

Natural Sciences I (Großer Hörsaal (Building 10.50))

- O. Koepler; C. Steinbeck; F. Bach; S. Herres-Pawlis; N. Jung; J. Liermann; S. Neumann; M. Razum
Digitalizing the Chemical Landscape: A Comprehensive Overview and Progress Report of NFDI4Chem
- L. Amelung; A. Barty; B. Murphy; C. Schneide; A. Schneidewind; T. Schoerner
The DAPHNE4NFDI and PUNCH4NFDI Consortia in the NFDI
- H. Weber; S. Brockhauser; C. Koch; L. Rettig; M. Aeschlimann; W. Hetaba; M. Grundmann; M. Kühbach; M. Krieger
Research Data Management for Experiments in Solid-State Physics: Concepts
- J. Bode; P. Jaeger; S. Schneidewind
Integrating Data Literacy into University Curricula: Student Centred Learning in Undergraduate Physics Lab Courses

Life Sciences (Kleiner Hörsaal (Building 10.50))

- J. Fluck; M. Golebiewski; J. Darms
Data publication for personalised health data: A new publication standard introduced by NFDI4Health
- Pigeot; J. Fluck; J. Darms; C. Schmidt
The NFDI4Health – Task Force COVID-19
- B. Ebert; J. Engel; I. Kostadinov; A. Güntsch; F. Glöckner
Connecting National and International Data Infrastructures in Biodiversity Research
- C. Goble; F. Bacall; S. Soiland-Reyes; S. Owen; I. Eguinoa; B. Droesbeke; H. Ménager; L. Rodriguez-Navas; J. Fernández; B. Grüning; S. Leo; L. Pireddu; M. Crusoe; J. Gustafsson; S. Capella-Gutierrez; F. Coppens
The EOSC-Life Workflow Collaboratory for the Life Sciences

Late afternoon (16:00 – 17:30h)

Humanities & Social Sciences II (Audimax (Building 30.95))

- S. Schneider; L. Palm
Sociodemographic variables in surveys: increasing research potential through output harmonization
- S. Netscher; A. Meyermann; J. Künstler-Sment; L. Pegelow
Stamp – Standardized Data Management Plan for Educational Research: An Approach to Improve Cross-Disciplinary Harmonization of Research Data Management
- P. Siegers; A. May; C. Saalbach; J. Nebelin; D. Kern; A. Daniel; B. Zapilko; F. Momeni; K. Wenzig; J. Goebel
Linked Open Research Data for Social Science: A concept registry for granular data documentation
- T. Emery; K. Karpinska; A. Maineri; L. van der Meer
The Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI): Better Infrastructure, Better Science, Better Society

Natural Sciences II (Großer Hörsaal (Building 10.50))

- R. Danabalan; M. Hintermüller; T. Koprucki; K. Tabelow
MaRDI: Building Research Data Infrastructures for Mathematics and the Mathematical Science
- P. Veluvali; J. Heiland; P. Benner
MaRDIFlow: A Workflow Framework for Documentation and Integration of FAIR Computation
- O. Knodel; T. Gruber; J. Kelling; M. Lokamani; S. Müller; D. Pape; M. Voigt; G. Juckeland
Overarching Data Management Ecosystem at HZDR: From Small Experiments to Large-Scale Research Facilities
- T. Gruber; H.-P. Schlenvoigt; O. Knodel; K. Tippey; G. Juckeland
Two-Step Approach in Metadata Management for Data Publications at Research Centers

Engineering Sciences (Kleiner Hörsaal (Building 10.50))

- R. Chacko; H. Goßler; J. Riedel; S. Schunk; O. Deutschmann
Digitalization in Catalysis and Reaction Engineering: Automating Work Flow

- P. Ost; Y. Shakeel; P. Tögel
Data Collections Explorer: An easy-to-use tool for sharing and discovering research data
- R. El-Athman; J. Rädler; O.Löhmman; A. Ariza; T. Muth
The BAM Data Store: Piloting an openBIS-Based Research Data Infrastructure in Materials Science
- O. Werth; S. Ferez; A. Niese; R. German; L. Huelk; C. Weinhardt; B. Vogel
Current Insights from Task Area 1 in NFDI4Energy: Building and Serving the Energy Research Community

Wednesday, 13.09.2023

08:00-09:00h	Registration Audimax Foyer (Building 30.95)			
09:00-10:30h	Poster Session II (Posters via Call for Papers) More information Audimax Foyer (Building 30.95)			
10:30-11:00h	Coffee break Audimax Foyer (Building 30.95)			
11:00-12:30h	Enabling RDM I Audimax (Building 30.95)	Harmonizing RDM I Großer Hörsaal (Building 10.50)	Connecting RDM I Kleiner Hörsaal (Building 10.50)	Securing RDM Seminarraum Forum A und B (Building 30.95)
12:30-14:00h	Lunch Market of the consortia Audimax Foyer (Building 30.95)			
14:00-15:30h	Enabling RDM II Audimax (Building 30.95)	Harmonizing RDM II Großer Hörsaal (Building 10.50)	Connecting RDM II Kleiner Hörsaal (Building 10.50)	Research Software Engineers birds-of-a-feather networking meeting (organised with de-RSE e.V.)

				Seminarraum Forum A und B (Building 30.95)
15:30- 16:00h	<p>Coffee break</p> <p>Market of the consortia</p> <p>Audimax Foyer (Building 30.95)</p>			
16:00- 17:30h	<p>Final discussion / Town hall meeting</p> <p>DFG: Panel discussion:</p> <p>„Die mit den Daten tanzen“! Über die Zukunft des Datenökosystems</p> <ul style="list-style-type: none"> • Research: <u>Kora Kristof</u> (Vizepräsidentin Digitalisierung und Nachhaltigkeit, KIT) • Public Sector: <u>Hanna Brenzel</u> (Leiterin Referat „Methoden der Datenanalyse“, Statistisches Bundesamt) • Industry: <u>Paul Heinz</u> (Head of R&D Digital Processes, Covestro) • Politics: <u>Alexander Dürnagel</u> (Leiter des Referats „Künstliche Intelligenz in der Mobilität, digitale Innovationen“, Bundesministerium für Verkehr und digitale Infrastruktur) • Moderation: <u>Carmen Hentschel</u> <p>More Information</p> <p>Audimax (Building 30.95)</p>			
17:30- 18:00h	Break			
18:00- 21:00h	Dinner Festsaal (Adenauerring 7) KIT	Excursion to KIT Campus North Meetingpoint: Outside of Audimax (Building 30.95)	Visit to the <u>Schlosslichtspiele</u> ("Castle Light Festival") Meeting point: Audimax Foyer (Building 30.95)	

* The event will be held in German.

Session Details

Morning (11:00 – 12:30)

Enabling RDM I (Audimax (Building 30.95))

- M Politze; I. Lang
coscine.nrw Landesweite Basisversorgung zur Verwaltung von Forschungsdaten im Open Source Modell
- F. Meineke; M. Golebiewski; X. Hu; T. Kirsten; M. Löbe; S. Klammt; U. Sax; W. Müller
NFDI4Health Local Data Hubs for Finding and Accessing Health Data: Making Distributed Data Accessible through a SEEK-Based Platform
- N. Fatima; P. Alper; F. Bianchini; K. Bösl; U. Wittig; C. Goble; F. Coppens
RDMkit: The Research Data Management Toolkit for Life Sciences
- E. Borisova; R. Abu Ahmad; G. Rehm
Open Science Best Practices in Data Science and Artificial Intelligence

Harmonizing RDM I (Großer Hörsaal (Building 10.50))

- B. Schembera; F. Wübbeling; T. Koprucki; C. Biedinger; M. Reidelbach; B. Schmidt; D. Göddeke; J. Fiedler
Building Ontologies and Knowledge Graphs for Mathematics and its Applications
- D. Iglezakis; D. Terzijska; S. Arndt; S. Leimer; J. Hickmann; M. Fuhrmans; G. Lanza
Modelling Scientific Processes with the m4i Ontology
- L. Castro; J. Fluck; D. Arend; M. Lange; D. Martini; S. Neumann; S. Schimmler; D. Rebholz-Schuhmann
Schema.org as a Lightweight Harmonization Approach for NFDI
- A. Behr; H. Borgelt; T. Petrenko; M. Dörr; N. Kockmann
Investigating the Landscape of Ontologies for Catalysis Research Data Management

Connecting RDM I (Kleiner Hörsaal (Building 10.50))

- L. Gadelha; J. Eufinger
The German Human Genome-Phenome Archive in an International Context: Toward a Federated Infrastructure for Managing and Analyzing Genomics and Health Data
- D. Müller; M. Umkehrer
International Data Access Network (IDAN) for sensitive microdata in Humanities & Social Sciences
- J. Bicarregui; S. Coles; B. Matthews; J. Frey; B. Montanari; V. Bunakov; N. Knight
Connecting Infrastructures: The Physical Sciences Data Infrastructure (PSDI) in the UK
- N. Weisweiler; R. Bertelmann; S. Genderjahn; H. Pampel
Connecting the Dots: The Helmholtz Research Data Ecosystem and its links to the NFDI

Securing RDM (Seminarraum Forum A und B (Building 30.95))

- E. Apondo; A. Züger; A. Bruns; K. Mehlis; C. Schickhardt; E. Winkler
Establishing Adaptive Governance in NFDI Consortia
- A. Bruns; S. Parker; F. Molnár-Gábor; E. Winkler
Developing Consent Tools for the Research Community at the German Human Genome-Phenome Archiv (GHGA)
- Yongli Mou; Feifei Li; Sven Weber; Sabith Haneef; Hans Meine; Liliana Caldeira; Mehrshad Jaberansary; Sascha Welten; Yeliz Yediel Ucer; Guido Prause; Stefan Decker; Oya Beyan; Toralf Kirsten

Distributed Privacy-Preserving Data Analysis in NFDI4Health with the Personal Health Train

- F. Boehm; U. Sax; O. Vettermann; P. Kamocki; V. Stoilova
„Hello ELSA, how are you?“

Afternoon (14:00 – 15:30)

Enabling RDM II (Audimax (Building 30.95))

- C. Beilschmidt; D. Brandenstein; J. Drönner; N. Glombiewski; M. Mattig; B. Seeger
On the Design and Implementation of Easy Access to External Spatiotemporal Datasets in NFDI
- M. Dieckmann; S. Beyvers; J. Hochmuth; A. Rehm; F. Förster; A. Goesmann
The Aruna Object Storage: A distributed multi cloud object storage system for scientific data management
- R. Macneil; T. Russell
RSpace + iRODS: A scalable, flexible and versatile solution that facilitates data and metadata interoperability and is suitable for deployment in conjunction with a wide range of e-infrastructures and Research Commons
- T. Zastrow; N. Fabas
Research Data Publication at Large Scale

Harmonizing RDM II (Großer Hörsaal (Building 10.50))

- U. Sax; C. Henke; C. Draeger; T. Bender; A. Kuntz; M. Golebiewski; H. Ulrich; M. Löbe
The Provenance Core Data Set: A Minimal Information Model for Data Provenance in Biomedical Research
- A. Wein; J. Reinkensmeier; A. Weidlich; J. Lilliestam; V. Hagenmeyer; M. Richter; S. Auer; A. Nieße; S. Lehnhoff
FAIR Data for Energy System Research: An Overview of NFDI4Energy Task Area 4
- G. Lanza; M. Koval; J.-L. Hippolyte; M. Iturrate-Garcia; O. Pellegrino; A.-S. Piette; F. Toro
Towards FAIR Research Data in Metrology
- M. Scheidgen; S. Brückner; S. Brockhauser; L. Ghiringhelli; F. Dietrich; A. Mansour; M. Albrecht; H. Weber; S. Botti; M. Aeschlimann; C. Draxl
FAIR research data with NOMAD: FAIRmat's distributed, schema-based research-data infrastructure to harmonize RDM in materials science

Connecting RDM II (Kleiner Hörsaal (Building 10.50))

- P. Wittenburg; U. Schwardmann; C. Bianchi; C. Weiland
FDOs to enable Cross-Silo Work
- M. Politze; Y. Shakeel; S. Hunke; P. Ost; R. Aversa; B. Heinrichs; I. Lang
Long Term Interoperability of Distributed Research Data Infrastructures
- O. Brand; V. Broda; M. Cyra; M. Fingerhuth; R. Gerlach; L. Gertis; B. Jacob; R. Müller-Pfefferkorn; H. Neuroth; S. Rehwald; J. Straka; B. Weiner
The Federal State Initiatives for RDM as intermediaries in a dynamic landscape of RDM infrastructures and services
- D. Fuß; M.-C. Laible
Data Trustees – They Do Work! The Example of Research Data Centers

Thursday, 14.09.2023

08:00-09:00h	Registration Audimax Foyer (Building 30.95)			
09:00-10:00h	Keynote: Mark Wilkinson , Universidad Politécnica de Madrid “A Series of FAIR Vignettes” More Information Audimax (Building 30.95)			
10:00-10:30h	Coffee break Audimax Foyer (Building 30.95)			
10:30-12:00h	Harmonizing RDM III Audimax (Building 30.95)	Enabling RDM III Großer Hörsaal (Building 10.50)	Spreading RDM I Kleiner Hörsaal (Building 10.50)	Linking RDM I Seminarraum Forum A und B (Building 30.95)
12:00-13:30h	Lunch Market of the consortia Audimax Foyer (Building 30.95)			
13:30-15:00h	Harmonizing RDM IV Audimax (Building 30.95)	Enabling RDM IV Großer Hörsaal (Building 10.50)	Spreading RDM II Kleiner Hörsaal (Building 10.50)	Linking RDM II Seminarraum Forum A und B (Building 30.95)
15:00-15:30h	Coffee break Market of the consortia Audimax Foyer (Building 30.95)			

15:30-17:00h	Rapporteur Talks & Closing Audimax (Building 30.95)
---------------------	---

Session details

Morning (10:00 – 12:00)

Harmonizing RDM III (Audimax (Building 30.95))

- S. Hagemann-Wilholt; A. Schrader; A. Czerniak
Isn't a number and a URL enough? Why PIDs matter and technical solutions alone are not sufficient.
- R. Baum; O. Koepler
Leveraging Terminology Services for FAIR Semantic Data Integration
- R. Huber; N. Karam; O. Koepler; P. Strömert
Finding a Common Ground for NFDI Terminologies: Proposing I-ADOPT as a NFDI Wide Semantic Layer
- M. Schröder; S. Genehr; R. Köhling; S. Schmidt; R. Schneider; S. Spors; G. Szepannek; D. Waltemath; F. Krüger
A survey on the current status of Research Data Management in Mecklenburg-Vorpommern: Preliminary results for a questionnaire study among researchers

Enabling RDM III (Großer Hörsaal (Building 10.50))

- L. Kulla; J. Bröder; C. Curdt; M. Kubin; H. Kollai; C. Lemster; M. Nolden; K. Schmieder; A. Strupp; K.-U. Stucky; E. Söding; K. Pascal Walter; A. Witold
The HMC Information Portal for enhanced metadata collaboration in the Helmholtz FAIR data space
- F. Henninger
Born-fair data projects using cookiecutter templates
- S. Schimmler; R. Altenhöner; L. Bernard; J. Fluck; A. Klinger; S. Lorenz; B. Mathiak; B. Miller; R. Ritz; T. Schörner-Sadenius; A. Sczyrba; R. Stein
Base4NFDI – Basic Services for NFDI: Creating NFDI-wide basic services in a world of specific domains
- M. Diepenbroek; I. Kostadinov; B. Seeger; F. Glöckner; M. Dieckmann; A. Goesmann; B. Ebert; S. Schimmler; Y. Sure-Vetter
Towards a Research Data Commons in the German National Research Data Infrastructure NFDI: Vision, Governance, Architecture

Spreading RDM I (Kleiner Hörsaal (Building 10.50))

- S. Leimer; S. Hendriks; L. Korte; J. Stegemann; S. Stock; H. Timm; S. Rehwald
Research Data Management Curriculum of the Research Data Services at the University Library Duisburg-Essen
- M. Richter; J. Putzke; T. Schimmer; A. Mehler-Bicher
We are still here, too! Research Data Management at Universities of Applied Sciences: Approaches from the Project „FDM@HAW.rlp“ in the German State Rhineland-Palatinate

- A. Erxleben-Eggenhofer; B. Batut
FAIR and scalable education: The Galaxy training network (GTN) and a Training Infrastructure as a Service (TlaaS)
- B. Slowig; M. Blümm; K. Förstner; B. Lindstädt; R. Müller; M. Lanczek
Der Zertifikatskurs „Forschungsdatenmanagement“ als Blaupause für die FDM-bezogene Kompetenzentwicklung im Rahmen der NFDI

Linking RDM I (Seminarraum Forum A und B (Building 30.95))

- F. Alshawaf; R. Guescini; F. Kotschka; M. Bierwirth; M. Dreyer
Harmonized research information for classifying and linking research data
- L. Rossenova; M. Schubotz; R. Shigapov
The case for a common, reusable Knowledge Graph Infrastructure for NFDI
- S. Auer; M. Stocker; O. Karras; A. Oelen; J. D'Souza; A.-L. Lorenz
Organizing Scholarly Knowledge in the Open Research Knowledge Graph
- H. Sack; T. Schrade; O. Bruns; E. Posthumus; T. Tietz; E. Norouzi; J. Waitelonis; H. Fliegl; L. Söhn; J. Tolksdorf; J. Steller; A. Azócar Guzmán; S. Fathalla; A. Ihsan; V. Hofmann; S. Sandfeld; F. Fritzen; A. Laadhar; S. Schimmler; P. Mutschke
Knowledge Graph based RDM Solutions: NFDI4Culture – NFDI-MatWerk – NFDI4DataScience

Afternoon (13:30 – 15:00)

Harmonizing RDM IV (Audimax (Building 30.95))

- O. Giraldo; D. Dessi; S. Dietze; D. Rebholz-Schuhmann; L. Castro
Machine-Actionable Metadata for Software and Software Management
- B. Heinrichs; M. Yazdi
Determining the Similarity of Research Data by Using an Interoperable Metadata Extraction Method
- M. Moser; J. Werheid; T. Hamann; A. Abdelrazeq; R. Schmitt
Which FAIR Are You? A Detailed Comparison of Existing FAIR Metrics in the Context of Research Data Management

Enabling RDM IV (Großer Hörsaal (Building 10.50))

- S. Schaaf; A. Erxleben-Eggenhofer; B. Grüning
Galaxy and RDM: Being more than a workflow manager: living the data life cycle
- F. Bach; K. Soltau; S. Göller; C. Bonatto Minella; S. Hofmann
RADAR: building a FAIR and community tailored Research Data Repository
- Y. Minamiyama; M. Hayashi; I. Fujiwara; J. Onami; S. Yokoyama; Y. Komiyama; K. Yamaji
Toward the development of NII RDC application profile using ontology technology
- P. Dolcet; M. Schulte; F. Maurer; N. Jung; R. Chacko; O. Deutschmann; J.-D. Grunwaldt
LabIMotion Electronic Lab Notebook as Research Data Management tool in Catalysis
- M. Doerr; S. Maak; M. Menke; U. Bornscheuer
The RDM System LARA: – semantics through automation from bottom up

Spreading RDM II (Kleiner Hörsaal (Building 10.50))

- C. van Gelder; A. Cardona; B. Leskošek; P. Palagi
Building Research Data Management (RDM) expertise and training resources in ELIXIR Nodes

- J. Ortmeier; F. Fink; A. Hoffmann; S. Herres-Pawlis
RDM in Chemistry: How to Educate and Train Future Researchers to Manage Their Data
- D. Waltemath; E. Inau; V. Satagopam; I. Balaur
Experiences from FAIRifying community data and FAIR infrastructure in biomedical research domains
- K. Behrens; K. Blask
RDM Compass: Building Competencies for the Professional Curation of Research Data

Linking RDM II (Seminarraum Forum A und B (Building 30.95))

- C. Speck; P. Jaquart; C. Weinhardt; J. Lilliestam; M. Schäfer; A. Weidlich; J. Zilles; N. Kerker
Transparency and Involvement of Society and Policy in a Data Sharing Platform
- R. Voshage; S. Sikder; S. Della Chiesa; T. Krüger; M. Schorcht; G. Meinel
Data, Tools and Services for spatial sustainability Science: The Story of the new IOER Research Data Centre
- M. Schäfer; R. Qussous; L. Hülk; J. Lilliestam; A. Weidlich
NFDI4Energy Case-Study: Comparative Analysis and Visualisation of Long-Term Energy System Scenarios
- A. Czech; V. Geenen; C. Breß; M. Turkovic Popovski; P. Krauß; T. Riedel; F. Gauterin
Designing a Mobility Data Trustee (MDT): Findings from a Multi-Disciplinary Analysis of Requirements of an MDT

Location: Karlsruher Institut für Technologie (KIT)

The “Audimax Foyer (Building 30.95)”, Straße am Forum 1, is the entrance area of the Audimax (Building 30.95). Registration, lunch and coffee breaks with catering take place here, as well as other activities. The “Seminarraum Forum A und B” is also in this building.

The rooms “Großer Hörsaal” and “Kleiner Hörsaal” are in building 10.50, Reinhard-Baumeister-Platz 1 (see map).

The „Festsaal“ is in the “Studentenhaus”, Adenauerring 7.

Volume 1

Proceedings of the Conference on Research Data Infrastructure 2023

“Connecting Communities”

12 – 14 September 2023, Karlsruhe

Connecting RDM

Gadelha and Eufinger	The German Human Genome-Phenome Archive in an International Context: Toward a Federated Infrastructure for Managing and Analyzing Genomics and Health Data	CoRDI2023-1
Müller and Umkehrer	International Data Access Network (IDAN) for Sensitive Microdata in Humanities & Social Sciences	CoRDI2023-2
Bicarregui et al.	Connecting Infrastructures: The Physical Sciences Data Infrastructure (PSDI) in the UK	CoRDI2023-3
Weisweiler et al.	Connecting the Dots: The Helmholtz Research Data Ecosystem and its Links to the NFDI	CoRDI2023-4
Wittenburg et al.	FDOs to Enable Cross-Silo Work	CoRDI2023-5
Politze et al.	Long Term Interoperability of Distributed Research Data Infrastructures	CoRDI2023-6
Brand et al.	The Federal State Initiatives for RDM as Intermediaries in a Dynamic Landscape of RDM Infrastructures and Services	CoRDI2023-7
Fuß and Laible	Data Trustees – They Do Work! The Example of Research Data Centers	CoRDI2023-8

Enabling RDM

Politze et al.	Coscine.nrw Landesweite Basisversorgung zur Verwaltung von Forschungsdaten im Open Source Modell	CoRDI2023-9
Meineke et al.	NFDI4Health Local Data Hubs for Finding and Accessing Health Data: Making Distributed Data Accessible Through a SEEK-Based Platform	CoRDI2023-10
Fatima et al.	RDMkit: The Research Data Management Toolkit for Life Sciences	CoRDI2023-11
Borisova et al.	Open Science Best Practices in Data Science and Artificial Intelligence	CoRDI2023-12

Beilschmidt et al.	On the Design and Implementation of Easy Access to External Spatiotemporal Datasets in NFDI	CoRDI2023-13
Dieckmann et al.	The Aruna Object Storage: A Distributed Multi Cloud Object Storage System for Scientific Data Management	CoRDI2023-14
Macneil and Russell	RSpace + iRODS: A Scalable, Flexible and Versatile Solution That Facilitates Data and Metadata Interoperability and is Suitable for Deployment in Conjunction With a Wide Range of E-infrastructures and Research Commons	CoRDI2023-15
Zastrow and Fabas	Research Data Publication at Large Scale	CoRDI2023-16
Kulla et al.	The HMC Information Portal for Enhanced Metadata Collaboration in the Helmholtz FAIR Data Space	CoRDI2023-17
Henninger	Born-fair Data Projects Using Cookiecutter Templates	CoRDI2023-18
Schimmler et al.	Base4NFDI - Basic Services for NFDI: Creating NFDI-wide Basic Services in a World of Specific Domains	CoRDI2023-19
Diepenbroek et al.	Towards a Research Data Commons in the German National Research Data Infrastructure NFDI: Vision, Governance, Architecture	CoRDI2023-20
Schaaf et al.	Galaxy and RDM: Being More Than a Workflow Manager: Living the Data Life Cycle	CoRDI2023-21
Bach et al.	RADAR: Building a FAIR and Community Tailored Research Data Repository	CoRDI2023-22
Minamiyama et al.	Toward the Development of NII RDC Application Profile Using Ontology Technology	CoRDI2023-23
Dolcet et al.	LabIMotion Electronic Lab Notebook as Research Data Management Tool in Catalysis	CoRDI2023-24
Engineering Sciences		
Chacko et al.	Digitalization in Catalysis and Reaction Engineering: Automatizing Work Flows	CoRDI2023-25
Ost et al.	Data Collections Explorer: An Easy-to-Use Tool for Sharing and Discovering Research Data	CoRDI2023-26
El-Athman et al.	The BAM Data Store: Piloting an OpenBIS-Based Research Data Infrastructure in Materials Science and Engineering	CoRDI2023-27

Werth et al.	Current Insights From Task Area 1 in NFDI4Energy: Building and Serving the Energy Research Community	CoRDI2023-28
--------------	--	--------------

Harmonizing RDM

Schembera et al.	Building Ontologies and Knowledge Graphs for Mathematics and its Applications	CoRDI2023-29
------------------	---	--------------

Iglezakis et al.	Modelling Scientific Processes With the m4i Ontology	CoRDI2023-30
------------------	--	--------------

Castro et al.	Schema.org as a Lightweight Harmonization Approach for NFDI	CoRDI2023-31
---------------	---	--------------

Behr et al.	Investigating the Landscape of Ontologies for Catalysis Research Data Management	CoRDI2023-32
-------------	--	--------------

Sax et al.	Provenance Core Data: Set A Minimal Information Model for Data Provenance in Biomedical Research	CoRDI2023-33
------------	--	--------------

Wein et al.	FAIR Data for Energy System Research: An Overview of NFDI4Energy Task Area 4	CoRDI2023-34
-------------	--	--------------

Lanza et al.	Towards FAIR Research Data in Metrology	CoRDI2023-35
--------------	---	--------------

Scheidgen et al.	FAIR Research Data With NOMAD: FAIRmat's Distributed, Schema-based Research-data Infrastructure to Harmonize RDM in Materials Science	CoRDI2023-36
------------------	---	--------------

Hagemann-Wilholt et al.	Isn't a Number and a URL Enough? Why PIDs Matter and Technical Solutions Alone are not Sufficient.	CoRDI2023-37
-------------------------	--	--------------

Baum and Koepler	Leveraging Terminology Services for FAIR Semantic Data Integration Across NFDI Domains: How to Integrate Terminology Services Into Other Service Applications	CoRDI2023-38
------------------	---	--------------

Huber et al.	Finding a Common Ground for NFDI Terminologies Proposing I-ADOPT as a NFDI Wide Semantic Layer	CoRDI2023-39
--------------	--	--------------

Schröder et al.	A Survey on the Current Status of Research Data Management in Mecklenburg-Vorpommern Preliminary Results for a Questionnaire Study Among Researchers	CoRDI2023-40
-----------------	--	--------------

Giraldo et al.	Machine-Actionable Metadata for Software and Software Management Plans for NFDI	CoRDI2023-41
----------------	---	--------------

Heinrichs and Yazdi	Determining the Similarity of Research Data by Using an Interoperable Metadata Extraction Method	CoRDI2023-42
---------------------	--	--------------

Moser et al.	Which FAIR are you? A Detailed Comparison of Existing FAIR Metrics in the Context of Research Data Management	CoRDI2023-43
--------------	---	--------------

Doerr et al.	The RDM System LARA - Semantics Through Automation From Bottom up	CoRDI2023-44
--------------	---	--------------

Humanities & Social Sciences

Kamocki et al.	Open Science and Language Data: Expectations vs. Reality: The Role of Research Data Infrastructures	CoRDI2023-45
----------------	---	--------------

Thiery et al.	Object-Related Research Data Workflows Within NFDI4Objects and Beyond	CoRDI2023-46
---------------	---	--------------

Fichtner et al.	WissKI: A Virtual Research Environment Based on Drupal	CoRDI2023-47
-----------------	--	--------------

Lieber et al.	MetaBelgica Project: A Linked Data Infrastructure Between Federal Scientific Institutes in Belgium	CoRDI2023-48
---------------	--	--------------

Schneider and Palm	Sociodemographic Variables in Surveys: Increasing Research Potential Through Output Harmonization	CoRDI2023-49
--------------------	---	--------------

Netscher et al.	Stamp - Standardized Data Management Plan for Educational Research: An Approach to Improve Cross-Disciplinary Harmonization of Research Data Management	CoRDI2023-50
-----------------	---	--------------

Siegers et al.	Linked Open Research Data for Social Science: A Concept Registry for Granular Data Documentation	CoRDI2023-51
----------------	--	--------------

Emery et al.	The Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI): Better Infrastructure, Better Science, Better Society	CoRDI2023-52
--------------	---	--------------

Life Sciences

Fluck et al.	Data Publication for Personalised Health Data: A New Publication Standard Introduced by NFDI4Health	CoRDI2023-53
--------------	---	--------------

Pigeot et al.	The NFDI4Health – Task Force COVID-19	CoRDI2023-54
---------------	---------------------------------------	--------------

Ebert et al.	Connecting National and International Data Infrastructures in Biodiversity Research: The Case of NFDI4Biodiversity, a German Consortium for Biodiversity, Ecology and Environmental Data	CoRDI2023-55
--------------	--	--------------

Goble et al.	The EOSC-Life Workflow Collaboratory for the Life Sciences	CoRDI2023-56
--------------	--	--------------

Linking RDM

Alshawaf et al.	Harmonized Research Information for Classifying and Linking Research Data	CoRDI2023-57
Rossenova et al.	The Case for a Common, Reusable Knowledge Graph Infrastructure for NFDI	CoRDI2023-58
Auer et al.	Organizing Scholarly Knowledge in the Open Research Knowledge Graph: An Open-Science Platform for FAIR Scholarly Knowledge	CoRDI2023-59
Sack et al.	Knowledge Graph Based RDM Solutions: NFDI4Culture - NFDI-MatWerk - NFDI4DataScience	CoRDI2023-60
Speck et al.	Transparency and Involvement of Society and Policy in a Data Sharing Platform	CoRDI2023-61
Voshage et al.	Data, Tools and Services for Spatial Sustainability Science: The Story of the New IOER Research Data Centre	CoRDI2023-62
Schäfer et al.	NFDI4Energy Case-Study: Comparative Analysis and Visualisation of Long-Term Energy System Scenarios	CoRDI2023-63
Czech et al.	Designing a Mobility Data Trustee (MDT): Findings From a Multi-Disciplinary Analysis of Requirements of an MDT	CoRDI2023-64

Natural Sciences

Koepler et al.	Digitalizing the Chemical Landscape: A Comprehensive Overview and Progress Report of NFDI4Chem	CoRDI2023-65
Amelung et al.	The DAPHNE4NFDI and PUNCH4NFDI Consortia in the NFDI	CoRDI2023-66
Weber et al.	Research Data Management for Experiments in Solid-State Physics: Concepts	CoRDI2023-67
Bode et al.	Integrating Data Literacy Into University Curricula: Student Centred Learning in Undergraduate Physics Lab Courses	CoRDI2023-68
Danabalan et al.	MaRDI: Building Research Data Infrastructures for Mathematics and the Mathematical Sciences	CoRDI2023-69
Veluvali et al.	MaRDIFlow: A Workflow Framework for Documentation and Integration of FAIR Computational Experiments	CoRDI2023-70

Knodel et al.	Overarching Data Management Ecosystem at HZDR: From Small Experiments to Large-Scale Research Facilities	CoRDI2023-71
Gruber et al.	Two-Step Approach in Metadata Management for Data Publications at Research Centers	CoRDI2023-72
Poster presentations II		
Neumann et al.	Harmonising, Harvesting, and Searching Metadata Across a Repository Federation	CoRDI2023-73
Kraft et al.	Terminologies in RDM for Engineering – a Service Approach: NFDI4Ing Terminology Service	CoRDI2023-74
Gu et al.	RDM Services at the Luxembourg National Data Service	CoRDI2023-75
Krasselt et al.	Swiss-AL: Platform for Language Data in Applied Sciences: On Challenges in the Field of Language Open Research Data	CoRDI2023-76
Reidelbach et al.	MaRDMO Plugin: Document and Retrieve Workflows Using the MaRDI Portal	CoRDI2023-77
Gaur et al.	Metadata Fields and Quality Criteria - XAS Reference Database under DAPHNE4NFDI	CoRDI2023-78
Klemm et al.	Opportunities and Limits of a Disciplinary Repository Using the Example MO RE data (eResearch Infrastructure for Motor Research Data)	CoRDI2023-79
Wissing et al.	Distributed Computing and Storage Infrastructure for PUNCH4NFDI	CoRDI2023-80
Brilhaus et al.	One Resource to Teach Them All	CoRDI2023-81
Pan et al.	Transparency and Involvement of the Energy-Related Industry in a Data Sharing Platform	CoRDI2023-82
Wiesing	WissKI Viewer: Casual Access for WissKI Data Sets	CoRDI2023-83
Rißler-Pipka et al.	Pathways Between National and European Research Infrastructures: A Humanities' Perspective	CoRDI2023-84
Saleh and Tochtermann	Architectural Design of BERD Information Portal	CoRDI2023-85
Rakers et al.	Because Data Shall Grow (and we With it): Steps Towards a Cultural Change for Sharing Research Data	CoRDI2023-86
Kiesler and Schiffner	Exploring and Improving Workflows for the Donation and Curation of Research Data	CoRDI2023-87

Moore and Kunis	Zarr: A Cloud-Optimized Storage for Interactive Access of Large Arrays	CoRDI2023-88
Bernard et al.	NFDI4Earth: Improving Research Data Management in the Earth System Sciences	CoRDI2023-89
Rache et al.	Digital Twin-Based Concept for Reliable Research Data Management: Integrating Proprietary Data Sources for Hyperspectral Imaging	CoRDI2023-90
Islam	Ten Simple Rules for Designing and Building a FAIR Research Infrastructure	CoRDI2023-91
Behr et al.	Ontology-Based Laboratory Data Acquisition With EnzymeML for Process Simulation of Biocatalytic Reactors	CoRDI2023-92
Mader and Kleemeyer	Castellum: A Data Protection-Compliant Web Application for the Subject Management of Human Science Studies	CoRDI2023-93
Herold et al.	FAIRification of Historical Geodata: Automated Metadata Extraction From Archival Maps	CoRDI2023-94
Schneidewind et al.	Several Partners – Joint Effort: RDM Synergies in Large Scale Research	CoRDI2023-95
Becker et al.	Data Management Plan Tools: Overview and Evaluation	CoRDI2023-96
Ebert et al.	When Data Crosses Borders – Join Forces! Multi-disciplinary Use Cases Within NFDI	CoRDI2023-97
Enke and Schörner-Sadenius	Science Data Platform and Digital Research Product	CoRDI2023-98
Berger et al.	CorWiz a Platform for Exploring Corrosion Data and Accessing Corrosion Models	CoRDI2023-99
Saldanha Bach et al.	FAIR Assessment Practices: Experiences From KonsortSWD and BERD@NFDI	CoRDI2023-100
Gonzalez-Marquez and Schmahl	Introducing JuOSC	CoRDI2023-101
Schörner et al.	PIDs in the Natural Sciences	CoRDI2023-102
Deckers and Rhiem	Exchanging Research Data with SampleDB	CoRDI2023-103
Svoboda et al.	The Data Steward Service Center (DSSC): FAIR-agro RDM-Expertise Hub	CoRDI2023-104
Dierkes et al.	Building the Next Generation of Data Savvy Biomedical Researchers	CoRDI2023-105

Bacher et al.	Spreading the Love for Mathematical Research Data	CoRDI2023-106
Hastik et al.	DALIA FAIR Open Educational Federation: Aggregation, Harmonisation, Curation, and Quality Assurance With DALIA	CoRDI2023-107
Wittenburg and Koureas	FDO to Structure the Domain of Knowledge	CoRDI2023-108
Mansour et al.	FAIRmat Guide to Writing Data Management Plans: A Practical Guide for the Condensed-Matter Physics and Materials-Science Communities	CoRDI2023-109
Hoffmann et al.	Embedding the de.NBI Cloud in the National Research Data Infrastructure Activities	CoRDI2023-110
Preuß et al.	Monitoring the State of Open and FAIR Data in Helmholtz: A Data-Harvesting and Dashboard-Approach by HMC	CoRDI2023-111
Usbeck et al.	NFDI4DS Gateway and Portal	CoRDI2023-112
Al Laban et al.	Establishing the Research Data Management Container in NFDIxCs	CoRDI2023-113
Castro et al.	RO-Crates Meets FAIR Digital Objects	CoRDI2023-114
Beccuti et al.	ICT Infrastructure Supporting the Italian Research Infrastructure on Microbial Resources MIRRI-IT	CoRDI2023-115
Boukhers and Castro	Enhancing Reproducibility in Research Through FAIR Digital Objects	CoRDI2023-116
Griem et al.	Automated Documentation of Research Processes Using RDM	CoRDI2023-117
von Suchodoletz et al.	Improving the Research Desktop Experience for OpenStack VDI: Integrating Hardware Accelerated Rendering and Remote Transport	CoRDI2023-118
von Suchodoletz et al.	DataPLANT Cloud Oriented Service Infrastructure: Open for Integration and Adaptation	CoRDI2023-119
Silva Pimenta et al.	A FAIR Future for Engineering Sciences: Linking an RDM Community Through a Scientific Journal	CoRDI2023-120
Zierop et al.	Conda, Container and Bots: How to Build and Maintain Tool Dependencies in Workflows and Training Materials	CoRDI2023-121
Kayikcioglu et al.	Interactive Tools (IT) in Galaxy: Combining Synchronous and Asynchronous Workflows	CoRDI2023-122
Wawer et al.	Quality Assessment for Research Data Management in Research Projects	CoRDI2023-123

Securing RDM

Apondo et al.	Establishing Adaptive Governance in NFDI Consortia: Lessons Learned from Deliberative Forums with Patients on their Role in the Governance of the German Human Genome-Phenome Archive (GHGA)	CoRDI2023-124
Bruns et al.	Developing Consent Tools for the Research Community at the German Human Genome-Phenome Archive (GHGA)	CoRDI2023-125
Mou et al.	Distributed Privacy-Preserving Data Analysis in NFDI4Health With the Personal Health Train	CoRDI2023-126
Boehm et al.	“Hello ELSA, how are you?”: Legal and Ethical Challenges in RDM, Current and Future Tasks of ELSA Activities Against the Background of AI and Anonymisation	CoRDI2023-127

Spreading RDM

Leimer et al.	Research Data Management Curriculum of the Research Data Services at the University Library Duisburg-Essen	CoRDI2023-128
Richter et al.	We are Still Here, too! Research Data Management at Universities of Applied Sciences: Approaches From the Project "FDM@HAW.rlp" in the German State Rhineland-Palatinate	CoRDI2023-129
Erleben-Eggenhofer et al.	FAIR and Scalable Education: The Galaxy Training Network (GTN) and a Training Infrastructure as a Service (TlaaS)	CoRDI2023-130
Slowig et al.	Der Zertifikatskurs „Forschungsdatenmanagement“ als Blaupause für die FDM-bezogene Kompetenzentwicklung im Rahmen der NFDI	CoRDI2023-131
van Gelder et al.	Building Research Data Management (RDM) Expertise and Training Resources in ELIXIR Nodes	CoRDI2023-132
Ortmeyer et al.	RDM in Chemistry: How to Educate and Train Future Researchers to Manage Their Data	CoRDI2023-133
Waltemath et al.	Experiences From FAIRifying Community Data and FAIR Infrastructure in Biomedical Research Domains	CoRDI2023-134
Behrens and Blask	RDM Compas: Building Competencies for the Professional Curation of Research Data	CoRDI2023-135

<https://www.doi.org/10.52825/cordi.v1i>

Editors

York Sure-Vetter, Nationale Forschungsdateninfrastruktur (NFDI) e.V. & Karlsruhe Institute of Technology (KIT)

Carole Goble, Information Management, University of Manchester

Review process

The authors submitted extended abstracts of up to 1000 words to the community tracks or the thematic tracks. These contributions were evaluated by the track chairs and then sent by them for external review to experts. The track chairs decided on required revisions and finally decided on their acceptance for the conference as presentation or poster in a second poster session and hence their publication in the proceedings.

Financing

The Proceedings of the CoRDI are financially supported by the Federal Ministry of Education and Research.

SPONSORED BY THE



Federal Ministry
of Education
and Research

The German Human Genome-Phenome Archive in an International Context: Toward a Federated Infrastructure for Managing and Analyzing Genomics and Health Data

Luiz Gadelha^[<https://orcid.org/0000-0002-8122-9522>] and Jan Eufinger^[<https://orcid.org/0000-0002-3439-1674>] on behalf of the GHGA-Consortium

German Human Genome-Phenome Archive, Germany

Abstract. With increasing numbers of human omics data, there is an urgent need for adequate resources for data sharing while also standardizing and harmonizing data processing. As part of the National Research Data Infrastructure (NFDI), the German Human Genome-Phenome Archive (GHGA) strives to connect the data from German researchers and their institutions to the international landscape of genome research. To achieve this, GHGA partners up with international activities such as the federated European Genome-Phenome Archive (EGA) [1] and the recently funded European Genomic Data Infrastructure (GDI) project to enable participation in international studies while ensuring at the same time the proper protection of the sensitive patient data included in GHGA.

Keywords: Genomics and Health Data, International Data Sharing, Federated Computing

1. Aims of GHGA and the need for international data sharing

To create a versatile and secure data infrastructure for genomics research, GHGA strives to provide (i) the necessary secure IT-infrastructure for Germany, (ii) an ethico-legal framework to handle omics data in a data-protection-compliant but open and FAIR [2] manner, (iii) harmonized metadata schemas, and (iv) standardized workflows to process the incoming omics data uniformly.

Genomic data is increasingly important in healthcare, allowing for clinical omics profiling of patients and enabling precision medicine, with tailored treatments having optimized efficiency for particular groups of patients. However, with many causal genetic alterations being typically very rare in given populations, successful knowledge generation in genome research critically depends on the availability of large cohorts of well curated reference datasets [3]. To exploit this potential it is therefore critical to share data at large scale, integrating data from different countries and their populations, incorporating diversity to the aggregated data set. To enable this, challenges bot on the technical side, caused by the need to manage and analyze the amounts of genomic data being collected, but also on the legal side, introduced by the inherent need for protection of individual's genome data, need to be met in a globally coordinated effort.

2. Federated Infrastructure to enable international science

Different projects are implementing federated infrastructures across national and continental scales in order to facilitate access to genomics and health data. The European Genome-Phenome Archive (EGA), run by the EMBL-EBI in the UK and the CRG in Spain, is one of the main repositories for genomics and health data started in 2008 [1]. There is ongoing work to expand EGA into a federated data infrastructure, the Federated EGA. In this model, countries will implement their own nodes following interoperability standards that will allow for sensitive data under controlled access to be stored in the national Federated EGA nodes. This avoids the need for further legal regulations and the exchange of person-related data with other nodes or the Central EGA (which corresponds to the original EGA). Similarly, the European Genomic Data Infrastructure (GDI) project is implementing a federated data infrastructure to allow access to genomics and health data across Europe. GDI will additionally allow for omics analysis workflows to be executed in the federated infrastructure. Jointly funded by the European Commission and the EU member states, GDI is an outcome of the Beyond 1 Million Genomes (B1MG) project and the 1+Million Genomes (1+MG) [4] initiative.

Positioned as the German national node both within the Federated EGA and the new GDI project, GHGA will enable German research projects to not only connect to those large scale international activities but also to contribute to the shaping of new standards and infrastructures for advancing research across Europe.

3. Global efforts to support genome research

Beyond these European activities, the Global Alliance for Genomics and Health (GA4GH) [5] provides many of the standards through which interoperability can be achieved in federated infrastructures. Consequently, both GHGA and its European partner projects include GA4GH developments into their mode of operation.

Data discovery is a key component of federated data infrastructures, and Beacon [6] is an example of data discovery protocol that enables researchers to search for and discover genomic and phenotypic data across different repositories and platforms without the need to expose sensitive information on individuals publicly. The Phenopackets [7] standard can be used for sharing phenotype and disease information along with associated data, such as genomics, diagnostics, and treatments. GA4GH Authentication and Authorization Infrastructure (AAI) and Passport standards [8] allow researchers to access data and resources across different platforms using a single set of credentials. The Data Use Ontology (DUO) [9] is used to annotate data sets with restrictions about their usage, standardizing the process of data access and use. Crypt4GH [10] provides a protocol for securely storing and sharing genomic data using public-key cryptography. GHGA closely follows and uses many of these standards to achieve interoperability for sharing genomics and health data within Germany and in Federated EGA and GDI. The GHGA Metadata Catalog was recently launched and consists of a public frontend for the discovery of study data from German research institutions, enabling the search of non-personal metadata. It aims to create a resource to collect information on human omics datasets available from German institutions for secondary research under controlled access conditions. The first data sets discoverable are 62 whole exomes, whole genomes and RNA sequencing data sets of 1310 patients suffering from 20 different rare cancer types from the NCT, DKFZ, and DKTK MASTER program. In the upcoming GHGA Archive, omics data will be stored and made available to other researchers after approval by the Data Access Committee of the data set controller.

GHGA is aiming to be more than an archive and consequently also contributes to the development and standardization of bioinformatics workflows for data analysis, benchmarking, statistical analysis, and visualizations. Here we are working with the global nf-core [11] community and also integrate GA4GH standards, such as the Workflow Execution Service (WES)

and Tool Execution Service (TES) to also enable the application of the FAIR principles [12] in workflow development.

By delivering a national IT infrastructure for data sharing and analysis, an ethico-legal framework, metadata schemas, and standardized and reproducible workflows, GHGA will enable cross-project analysis and promote new collaborations and research projects in the international context of genome research. This will also be of high importance for new national projects such as e.g. the upcoming genomDE project within Germany.

Author contributions

Luiz Gadelha and Jan Eufinger wrote the initial draft. Members of the GHGA-Consortium revised and supervised the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

The research project GHGA - German Human Genome-Phenome Archive is funded by the German Research Foundation (DFG) within the framework of the National Research Data Infrastructure (NFDI).

References

1. M. A. Freeberg *et al.*, "The European Genome-phenome Archive in 2021," *Nucleic Acids Res.*, vol. 50, no. D1, pp. D980–D987, Jan. 2022, doi: 10.1093/nar/gkab1059.
2. M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
3. Z. Stark *et al.*, "Integrating Genomics into Healthcare: A Global Responsibility," *Am. J. Hum. Genet.*, vol. 104, no. 1, pp. 13–20, Jan. 2019, doi: 10.1016/j.ajhg.2018.11.014.
4. G. Saunders *et al.*, "Leveraging European infrastructures to access 1 million human genomes by 2022," *Nat. Rev. Genet.*, vol. 20, no. 11, pp. 693–701, Nov. 2019, doi: 10.1038/s41576-019-0156-9.
5. H. L. Rehm *et al.*, "GA4GH: International policies and standards for data sharing across genomic research and healthcare," *Cell Genomics*, vol. 1, no. 2, p. 100029, Nov. 2021, doi: 10.1016/j.xgen.2021.100029.
6. J. Rambla *et al.*, "Beacon v2 and Beacon networks: A 'lingua franca' for federated data discovery in biomedical genomics, and beyond," *Hum. Mutat.*, p. humu.24369, Apr. 2022, doi: 10.1002/humu.24369.
7. J. O. B. Jacobsen *et al.*, "The GA4GH Phenopacket schema defines a computable representation of clinical data," *Nat. Biotechnol.*, vol. 40, no. 6, pp. 817–820, Jun. 2022, doi: 10.1038/s41587-022-01357-4.
8. C. Voisin *et al.*, "GA4GH Passport standard for digital identity and access permissions," *Cell Genomics*, vol. 1, no. 2, p. 100030, Nov. 2021, doi: 10.1016/j.xgen.2021.100030.
9. J. Lawson *et al.*, "The Data Use Ontology to streamline responsible access to human biomedical datasets," *Cell Genomics*, vol. 1, no. 2, p. 100028, Nov. 2021, doi: 10.1016/j.xgen.2021.100028.
10. A. Senf *et al.*, "Crypt4GH: a file format standard enabling native access to encrypted data," *Bioinformatics*, vol. 37, no. 17, pp. 2753–2754, Sep. 2021, doi: 10.1093/bioinformatics/btab087.

11. P. A. Ewels *et al.*, "The nf-core framework for community-curated bioinformatics pipelines," *Nat. Biotechnol.* 2020 383, vol. 38, no. 3, pp. 276–278, Feb. 2020, doi: 10.1038/s41587-020-0439-x.
12. C. Goble *et al.*, "FAIR Computational Workflows," *Data Intell.*, vol. 2, no. 1–2, pp. 108–121, Jan. 2020, doi: 10.1162/dint_a_00033.
13. M. Herschel, R. Diestelkämper, and H. Ben Lahmar, "A survey on provenance: What for? What form? What from?," *VLDB J.*, vol. 26, no. 6, pp. 881–906, Dec. 2017, doi: 10.1007/s00778-017-0486-1.
14. S. Cohen-Boulakia *et al.*, "Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 75, pp. 284–298, Oct. 2017, doi: 10.1016/j.future.2017.01.012.
15. J. Ison *et al.*, "EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats," *Bioinformatics*, vol. 29, no. 10, pp. 1325–1332, May 2013, doi: 10.1093/bioinformatics/btt113.
16. A. Gray, C. Goble, and R. Jimenez, "Bioschemas: From Potato Salad to Protein Annotation," in *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017)*, 2017.

International Data Access Network (IDAN) for sensitive microdata in Humanities & Social Sciences

Dana Müller¹, Matthias Umkehrer¹

¹ Institute for Employment Research, Germany

Abstract. In a globalized world it becomes increasingly important to provide international research data. This is particularly true in Humanities & Social Sciences. With legal frameworks changing, administrative data can increasingly be utilised both for official statistics and to facilitate new research, enabling the development of evidence-based policy for the public benefit. Secure access conditions generally apply to using these rich, highly detailed level data. However, using data from various sources is difficult when they are fragmented in ‘silos’ between several **Research Data Centres** (RDCs). While this might be the case at a national level, it is very likely to be the case at an international level. The latter is a major obstacle for international comparative research.

The International Data Access Network (IDAN, <https://idan.network/>) aims at creating a concrete operational international framework enabling access to controlled data for research. The Network was founded in 2018, at the same time when the European General Data Protection Regulation became active. It currently involves six RDCs from France, Germany, the Netherlands and the United Kingdom. Within IDAN, step by step cooperative solutions are developed taking into account the particular legal and security requirements that are at the core of both, national and transnational access to **confidential data**. Initially, the partners’ access systems are being implemented in each partners’ premise based on bilateral agreements. This process involves combining requirements of security and surveillance for Safe Rooms. It sets up a new concrete environment for researchers to work remotely with data from the other partners from their local RDC.

IDAN is a perfect match for the topic “Connecting RDM” regarding confidential data which are a highly demanded by the international research community. The presentation will show the opportunities and obstacles in developing the IDAN infrastructure. Based on discussions with research experts with heterogeneous backgrounds in Europe, it will also discuss researcher’s expectations and needs to further improve access to international research data. IDAN could be a role model for other research disciplines and research infrastructure for pushing the boundaries of international **data access**.

Keywords: Research Data Infrastructure

Connecting Infrastructures: The Physical Sciences Data Infrastructure (PSDI) in the UK

Juan Bicarregui¹[\[https://orcid.org/0000-0001-5250-7653\]](https://orcid.org/0000-0001-5250-7653), Simon J Coles²[\[https://orcid.org/0000-0001-8414-9272\]](https://orcid.org/0000-0001-8414-9272), Brian Matthews¹[\[https://orcid.org/0000-0002-3342-3160\]](https://orcid.org/0000-0002-3342-3160), Jeremy G Frey²[\[https://orcid.org/0000-0003-0842-4302\]](https://orcid.org/0000-0003-0842-4302), Barbara Montanari¹[\[https://orcid.org/0000-0001-8654-9181\]](https://orcid.org/0000-0001-8654-9181), Vasily Bunakov¹[\[https://orcid.org/0000-0003-3467-5690\]](https://orcid.org/0000-0003-3467-5690), and Nicola J Knight³[\[https://orcid.org/0000-0001-8286-3835\]](https://orcid.org/0000-0001-8286-3835)

¹ Scientific Computing Department, Science and Technologies Facilities Council, UK

² School of Chemistry, University of Southampton, UK

Abstract. In this presentation we discuss the activities undertaken in the UK through the Physical Sciences Data Infrastructure (PSDI) initiative, part of the wider Digital Research Infrastructure (DRI) Programme. We will present the aims of the PSDI initiative, our initial scoping work and trials, and suggest how this project can and should interact with other related initiatives on a national and global scale.

Physical Scientists are crying out for a socio-technical data infrastructure that connects existing experimental and computational facilities. We believe that a cross-discipline and cross-technique digital infrastructure that builds on and bridges across existing initiatives, while these continue to serve their particular fields, is crucial to, and the best way to achieve global collaborations in the 21st century.

Keywords: Infrastructure, Physical Science

1. Introduction

In this presentation we discuss the activities undertaken in the UK through the Physical Sciences Data Infrastructure (PSDI) initiative, part of the wider Digital Research Infrastructure (DRI) Programme. We will present the aims of the PSDI initiative, our initial scoping work and trials, and suggest how this project can and should interact with other related initiatives on a national and global scale.

The data needs of research are growing at previously unimaginable rates and the need for collaboration around data has never been clearer. Data is not simply an output of research but is itself a driver of further discovery. Experiments, observations, computations and simulations all generate data. From simple manual annotations to complex simulations and terabytes of measurements from bespoke equipment, data flows are the very fabric of research but are currently hampered by technical and social problems related to data discovery, access, integration, processing, curation and publication.

Physical Scientists are crying out for a socio-technical data infrastructure that connects existing experimental and computational facilities. We believe that a cross-discipline and cross-technique digital infrastructure that builds on and bridges across existing initiatives, while these continue to serve their particular fields, is crucial to, and the best way to achieve global collaborations in the 21st century.

The aim of PSDI is to enable researchers in the physical sciences to handle data more easily by connecting the different data infrastructures they use. PSDI will connect and enhance existing infrastructure in Physical Sciences.

Through PSDI researchers will be able to:

- Find and Access to reference quality data from commercial and open sources
- Combine data from different sources
- Share data, software and models including experimental and simulation data
- Use AI to explore data
- Learn how to make the results of their research open and FAIR

2. Statement of Need

The PSDI project was initially discussed as part of the EPSRC Large Infrastructure Investments Statement of Need (SoN) call, which was submitted in early 2021[1]. During this SoN exercise a project team from STFC and the University of Southampton developed the outline plan for the PSDI. This included commentary on the ambition of the project and the strategic importance of investment in infrastructure for the physical sciences. This SoN exercise was well supported across the physical sciences community. Contributions and backing from a wide range of projects and initiatives demonstrated a community need and support for such an initiative.

The SoN confirmed a widespread consensus in the community that investment in research data infrastructure is lagging behind investment in data sources and identified an urgent need for integration of data and computational infrastructures. It identified four major 'pillars' of user communities in the UK that would benefit from the proposed PSDI. (Figure 1)

- Pillar 1. Facilities, Institutes and Hubs – significant centralised national facilities and activities that serve many researchers based on a common need.
- Pillar 2. National Research Facilities – medium-scale centralised facilities operating at a world leading level to perform research that cannot be addressed in a standard laboratory.
- Pillar 3. Computational Initiatives – uniting performing simulations with the communities and tools required to do so.
- Pillar 4. Research Institutions, research groups and laboratories.

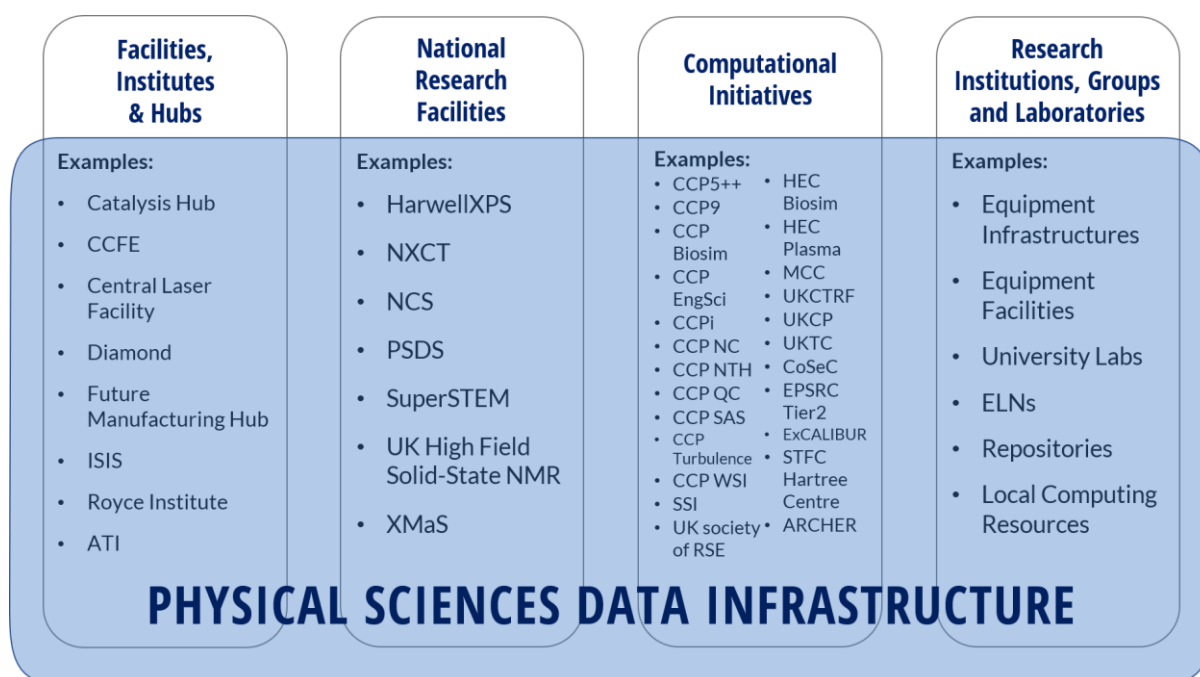


Figure 1 4 Pillars of UK User Communities producing and using data in UK Physical Science

3. PSDI Pilot

Following on from the SoN application, the PSDI team were requested by EPSRC to complete a proposal for a short phase pilot project, funded through the UKRI DRI programme. The PSDI is a large undertaking, involving a wide range of stakeholders within our proposed pillars and the wider community including data managers, data providers, system architects and many more roles that underpin our national data landscape. This pilot ran for 5 months in late 2021 – early 2022 and expanded on the ambitions of the project from the SoN. In this pilot PSDI undertook a wide range of community consultation on the scope and requirements of the PSDI, and planning for the future of PSDI. (Figure 2)



Figure 2 A summary of the PSDI Pilot phase activities

As a result of the PSDI pilot a series of reports [2,3] and recommendations [4] were published. There were 13 recommendations across 4 areas, summarised below:

- Connecting existing infrastructure: connecting existing research data services, support beyond the lifespan of individual projects, co-operation and co-creation between all stakeholder organisations
- Best Use of Data: developing a toolkit for publishing, access to provenanced data, tools for reproduceable data processing, support for transforming data to knowledge
- Best Use of People: co-ordination for community activities and input, community training and support, professionalisation for data roles, governance structure for PSDI
- Best Use of Technology: services to connect existing provision (data and services), adopt existing technologies

4. Current Phase

In 2023 PSDI has undertaken further scoping and prototype development work towards the realisation of the PSDI vision. As part of this phase 5 pathfinder activities are underway in the areas of: Catalysis, Process recording, Data collections, Bio-molecular simulations and Data to Knowledge.

As PSDI develops beyond into the future, interaction with other data initiatives across UKRI, and the wider international sphere, will continue to be a central focus. The components that drive science forward, such as data, standards, and research developments are not limited to a single research community or country but require connection and integration on a global scale.

The life sciences have had considerable financial input to create global infrastructure to support data sharing. Examples include the Protein Data Bank (PDB), Gene sequences, etc. The ability to access the protein structures was a key aspect in the success of DeepMind's development of AlphaFold2 and the need to share sequence information was key in the WHO's work in tracking the COVID-19 global pandemic. Particle Physics has developed a global computational infrastructure to support the specific types of data produced by organisations such as CERN. However, the wider, longer tail, smaller research group based Physical Sciences are ca, 20 years behind these efforts and PSDI together with other international initiatives seeks to make a start on bringing our community the benefits.

Data availability statement

The content of this presentation is not derived directly from specific datasets. However, all reports from the pilot phase are available through the PSDI zenodo community <https://zenodo.org/communities/psdi>

Underlying and related material

N/A

Author contributions

All authors were involved in conceptualization, funding acquisition and project administration. NJK prepared the original draft, and all authors contributed to the review & editing of the submission.

Competing interests

The authors declare that they have no competing interests.

Funding

PSDI is funded by EPSRC grants EP/X032701/1, EP/X032663/1 and EP/W032252/1

Acknowledgement

We acknowledge those members of the community who contributed to our Statement of Need exercise, who carried out work in the PSDI pilot (<https://www.psd.ac.uk/the-pilot/team/>), or current phase of our work (<https://www.psd.ac.uk/people/>).

References

1. "Physical Science Data Infrastructure Statement of Need." PSDI. https://www.psd.ac.uk/wp-content/uploads/2022/01/PSDI_SoN-V1.0_OneDocument_LargeInfrastructureInvestments.pdf (accessed: 21/04/2023)
2. N J Knight, J Bicarregui, B Montanari, S J Coles, J G Frey, B Matthews, & V Bunakov. (2023). Physical Sciences Data Infrastructure Phase 1 Pilot Report. Zenodo. <https://doi.org/10.5281/zenodo.7684860>
3. "Physical Sciences Data Infrastructure Community." Zenodo. <https://zenodo.org/communities/psdi> (accessed: 21/04/2023)
4. "Pilot Recommendations." PSDI. <https://www.psd.ac.uk/the-pilot/recommendations> (accessed 21/04/2023)

Connecting the Dots

The Helmholtz Research Data Ecosystem and its Links to the NFDI

Nina Leonie Weisweiler^{1,2}[\[https://orcid.org/0000-0001-6967-9443\]](https://orcid.org/0000-0001-6967-9443), Roland Bertelmann¹[\[https://orcid.org/0000-0002-5588-0290\]](https://orcid.org/0000-0002-5588-0290), Steffi Genderjahn¹[\[https://orcid.org/0000-0002-8912-184X\]](https://orcid.org/0000-0002-8912-184X), and Heinz Pampel^{1,3}[\[https://orcid.org/0000-0003-3334-2771\]](https://orcid.org/0000-0003-3334-2771)

¹ Helmholtz Association, Helmholtz Open Science Office

² Helmholtz Association, Helmholtz Metadata Collaboration

³ Humboldt-Universität zu Berlin

The Helmholtz Centers operate scientific-technical infrastructures that produce a high volume of digital research data, making Helmholtz a hub for expertise in research data. With the rapid digital transformation and growing volume of data, Helmholtz has implemented relevant policies and engaged in NFDI in various ways to manage and use research data effectively.

The Helmholtz Information & Data Science Incubator, in particular the HMC and HIFIS platforms, contribute to these networking activities. The Helmholtz Open Science Office organized several internal forums for Helmholtz members to discuss NFDI, their findings will be presented at the Conference on Research Data Infrastructure (CoRDI).

Helmholtz' multifaceted commitment to research data infrastructure is closely related to NFDI activities. The presentation at CoRDI will demonstrate how a large national research data ecosystem can be successfully connected to the NFDI network, highlighting opportunities for future collaboration.

Keywords: Helmholtz Association, Helmholtz Open Science Office, Helmholtz Information & Data Science Incubator, Networking

- Based on the Helmholtz mission, the 18 Helmholtz Centers operate complex scientific-technical infrastructures, such as particle accelerators, satellites or research ships and aircraft which produce a high volume of digital research data. Because of its data-intensive research practices, Helmholtz is a hub for expertise in research data, including aspects such as access to and reuse of research data.
- Due to the rapid digital transformation and the constantly growing volume of data, the requirements towards management and use of research data within and beyond Helmholtz are changing fundamentally. Helmholtz takes this into account through relevant policies in the Centers and a Helmholtz Open Science Policy [1].
- As the National Research Data Infrastructure (NFDI) builds a central infrastructure in this context, numerous NFDI consortia are implemented with significant Helmholtz participation. All Helmholtz Centers have become members of the NFDI e.V. association and several Helmholtz members are engaged in the NFDI Sections.

Beteiligung der Helmholtz-Zentren an der Nationalen Forschungsdateninfrastruktur (NFDI)



Figure 1. Participation of Helmholtz Centers in NFDI consortia and in the NFDI Association.

- In addition, the Helmholtz-internal platforms in the Helmholtz Information & Data Science Incubator, in particular HMC and HIFIS, have established strong data-relevant links between the Centers and contribute in many ways to the realization of NFDI, e.g., in the context of Base4NFDI.
- On an international level, Centers are members in EOSC-A and contribute to EOSC-relevant projects. Furthermore, there is a broad engagement in numerous international data-driven scientific projects for example in the context of the European Strategy Forum on Research Infrastructures (ESFRI).
- The Helmholtz Open Science Office promotes discussion about the reuse and reproducibility of research data in Helmholtz and offers a platform for interdisciplinary exchange in this field. To stimulate dialogue on the topic of NFDI in Helmholtz, the Helmholtz Open Science Office organized, among other activities, two forums for Helmholtz members in May and December 2021 (see the reports [2],[3]). A third forum is planned for June 2023. The findings from these events will be incorporated into the presentation at the Conference on Research Data Infrastructure (CoRDI).
- Our presentation at the CoRDI will show how Helmholtz structures interact with the NFDI, providing an example of how a large national research data ecosystem can be

successfully connected to the NFDI network. We will demonstrate how Helmholtz' multifaceted commitment to research data infrastructure is related to NFDI activities and what opportunities we see in this collaboration for the future.

Data availability statement

The submission is not based on data analysis.

Underlying and related material

Not applicable.

Author contributions

Nina Weisweiler: Writing – original draft

Roland Bertelmann, Steffi Genderjahn, and Heinz Pampel: Writing – review & editing

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Acknowledgement

Not applicable.

References

1. Helmholtz-Gemeinschaft (Ed.) (2022): Helmholtz Open Science Policy. Version 1.0. In der 119. Mitgliederversammlung der Helmholtz-Gemeinschaft am 20.-21. September 2022 beschlossen, Helmholtz-Gemeinschaft, doi: <https://doi.org/10.48440/os.helmholtz.055>
2. Weisweiler, N. L., Bertelmann, R., Braesicke, P., Bronger, T., Curdt, C., Glöckner, F. O., Rank, S., Stegle, O., Sure-Vetter, Y., Villacorta, N. (2021): Helmholtz Open Science Briefing: Helmholtz in der Nationalen Forschungsdateninfrastruktur (NFDI): Report des Helmholtz Open Science Forums, doi: <https://doi.org/10.48440/os.helmholtz.030>
3. Weisweiler, N. L., Bertelmann, R., Curdt, C., Glöckner, F. O., Jandt, U., Streit, A., Sure-Vetter, Y., Villacorta, N. (2022): Helmholtz Open Science Briefing: Zweites Helmholtz Open Science Forum „Helmholtz in der Nationalen Forschungsdateninfrastruktur (NFDI)“: Report, doi: <https://doi.org/10.48440/os.helmholtz.037>

FDOs to enable Cross-Silo Work

Peter Wittenburg¹[\[https://orcid.org/0000-0003-3538-0106\]](https://orcid.org/0000-0003-3538-0106), Ulrich Schwardmann², Christoph Bianchi³,
Claus Weiland⁴

¹ FDO Forum, International

² FDO Forum, International

³ DONA Foundation, Switzerland

⁴ Senckenberg Institute, Germany

Abstract. In this paper we describe with the help of two examples how FAIR Digital Objects can be used to bridge between different dataspace, repositories using different technologies, and model worlds. The first example is the digital specimen as defined by DiSSCO which offers access to a variety of information from different repositories in a persistent way. The second example results from a collaboration with the I4.0 Asset Administration Shell experts where we integrate the two domains using the Digital Product Pass as an example. In this example we also show how FDOs can be used to implement secure access on shared data as it occurs in almost all supply chain processes.

Keywords: FAIR Digital Objects, Secure Data, Data Management, Data Interoperability

1. Introduction

Extensive funding programs especially in Europe (EOSC, NFDI, Gaia-X, etc.) are intensifying the work on data infrastructures in research and industry with the intention to create deep knowledge on all levels from policy to development to finally manage the digital transformation. In parallel the major information companies are also investing huge sums to provide proprietary infrastructure services in a highly competitive landscape. The motivation for these investments becomes increasingly more obvious and urgent: How to democratise the ability to apply statistical methods such as Deep Learning based on AI-ready data.

Each of the many projects being funded create their specific dataspace aligned with a specific technology stack and a set of rules. If we extend this increasing European fragmentation to the global level we can imagine that we will have an interoperability challenge at several levels – from data modelling to semantics. This situation can be compared with the lack of network technology interoperability a few decades ago when a wide variety of networking solutions were developed until a consensus was built around the internet technology: (1) The agreement on “datagrams” as autonomous entities travelling within the global network. (2) The agreement on TCP/IP as a unifying interface protocol.

2. FAIR Digital Objects

The FAIR Digital Object model (FDO, figure 1) is analogous to the datagrams as autonomous, self-contained and machine actionable entities that persistently bind all information necessary to facilitate its processing across a wide range of different dataspace independently of their

respective technological choices and sets of rules [1]. Therefore, FDOs are a candidate for achieving basic global data interoperability. The FDO's common data model emerged from

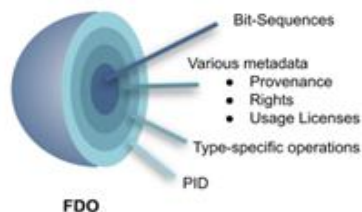


Figure 1. Schematic indication of the FDO Concept.

efforts started within the RDA covering multiple scientific disciplines. Over the last three years the major focus of the FDO work was on combining the basic object model with the FAIR principles and to make it machine actionable. As a result the FDO specifications require that each digital entity be assigned a globally unique, persistent identifier that is resolvable into a predictable, profile specified, set of attribute value pairs that are easily readable and interpretable by any client. Each attribute will be defined and registered in open registries [2]. These attributes can be used to specify a wide range of properties such as checksums to verify the integrity of data, or signatures to verify the authenticity of a claim, or references to resources (data, metadata, configurations, software services, etc.) hosted on repositories of measurable trustworthiness.



Figure 2. Possibility of creating an integrated FDO domain by using a unique protocol.

In the same way that datagrams are transported on the Internet using TCP/IP, a unifying protocol we call DOIP [3] provides a standard approach for interacting with any FDOs on the Internet independently of the specific technologies used by the various service providers. DOIP acts as the interoperability glue between data service providers of different sorts (figure 2). As an example, we can refer to the development of a DOIP adapter that allowed us to connect the B2Share repository [4] with the CORDRA repository [5] both using different technologies and data models. Other groups have also been using DOIP in a wide range of projects and at large scale. We are currently in the process of integrating some of these different projects in a comprehensive testbed and building on work from KIT and GWDG. This testbed will include repositories from different disciplines and countries/regions to demonstrate the usefulness and scalability of the approach. The goal is to provide validators to enable anyone to connect their repository.

3. Interconnecting with FDOs

The Digital Specimen (DS, figure 3) developed in the context of the DiSSCo RI demonstrates the huge potential of FDOs in managing digital artefacts which are created and curated by numerous institutions [6]. The DS combines a wide range of information of different types (images like scans, sequence data, geolocations, taxonomic classifications, etc.) acting as a Digital Twin of a physical specimen in a scientific collection. The DS concept makes also use of another feature of FDOs: direct links between type and registered operations. In DiSSCo, many

FDOs are accessed by clients using operations like feature extraction from digitised content. Such operations can be registered in a registry managed by the resource provider or in open registries managed by other communities of practice.

Intensive interactions with experts who designed and developed the Industry 4.0 Asset Administration Shell (AAS) solution resulted in the design of a smart solution for the common

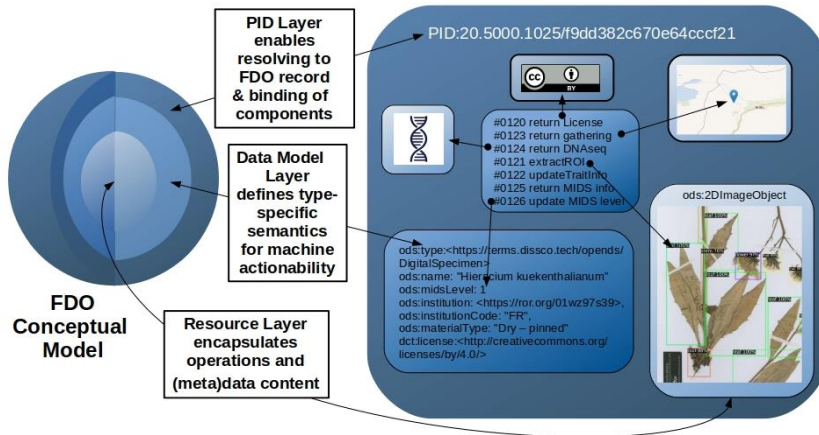


Figure 3. Illustration of the DiSSCO Digital Specimen concept.

challenge of enabling many distributed actors to operate on a shared data structure in a highly secure manner [7]. An example application motivated by industry is the development of an FDO based tracking of the transportation related green-gas emissions integrated with the Dig-

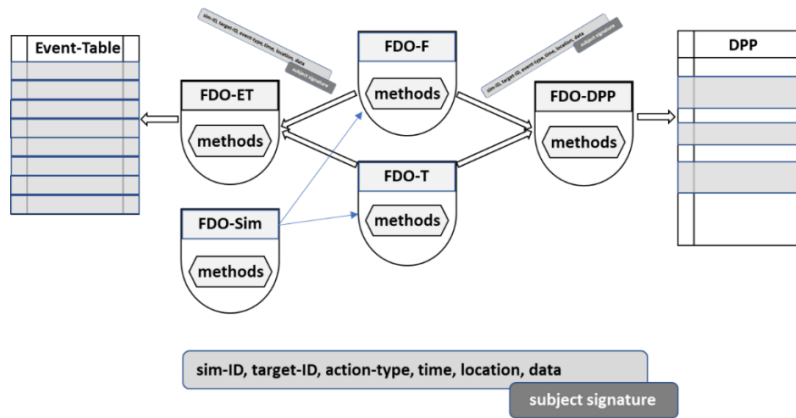


Figure 4. Concept to implement secure sharing of the Digital Product Pass which modelled using the I4.0 AAS technology.

ital Product Pass. All data structures and actors including the DPP are modelled as FDOs and the use of registered typed operation relations is used to enforce secure interactions (figure 4). Operations on FDOs can be compared to methods as they are known from object-oriented programming. Standard security technologies such as digital signatures certificates and PKI infrastructures are core to this approach. This solution demonstrates that industry standards such as I4.0 AAS and FDOs can be combined to achieve smart solutions working across country borders. Stakeholders from industry see much potential here to fill the gap between Operational Technology and Information Technology, which is a key issue for industry.

3. Conclusion

In this contribution based on concrete examples we show that FDOs

- due to their persistent and machine actionable bundling of information are a candidate for establishing the emerging Global Integrated Dataspace
- provide a simple yet extensible solution to structure the huge amount of distributed data sources around concepts such as the digital specimen and to provide this information in a persistent way
- can be used to implement secure interactions on shared resources in a smart way.

FDOs is an open, licence and property rights free concept involving a variety of implementations that will be turned into an International Standard with the help of DIN to enable broad application.

Author contributions

All authors contributed to the whole paper.

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

We would like to acknowledge the work of the FDO community and the excellent collaboration with our colleagues from RWTH Aachen who were involved in the specification and implementation of the I4.0 AAS concept.

References

1. I. Anders, et al., "FAIR Digital Object Technical Overview", <https://zenodo.org/record/7824714#.ZEDULs7P3b0>
2. I. Anders, et al., "FDO Forum FDO Requirement Specifications", <https://zenodo.org/record/7782262#.ZEDURs7P3b0>
3. DONA, "Digital Object Interface Protocol Specification", https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf
4. EUDAT. "B2SHARE", <https://b2share.eudat.eu/>
5. DONA, "CORDRA – DOIP Client Library", <https://www.cordra.org/documentation/client/doip-java.html>
6. DiSSCO, "What is a Digital Specimen?", <https://dissco.tech/2020/03/31/what-is-a-digital-specimen/>
7. P. Wittenburg, C. Bianchi, "Interacting FDOs for Secure Processes", https://www.researchgate.net/publication/364508074_Interacting_FDOs_for_Secure_Processes

Long Term Interoperability of Distributed Research Data Infrastructures

Marius Politze¹[\[https://orcid.org/0000-0003-3175-0659\]](https://orcid.org/0000-0003-3175-0659), Yusra Shakeel²[\[https://orcid.org/0000-0001-5135-4325\]](https://orcid.org/0000-0001-5135-4325),
Sirieam Hunke¹[\[https://orcid.org/0000-0001-9316-4220\]](https://orcid.org/0000-0001-9316-4220), Philipp Ost²[\[https://orcid.org/0000-0002-7198-0566\]](https://orcid.org/0000-0002-7198-0566),
Rossella Aversa²[\[https://orcid.org/0000-0003-2534-0063\]](https://orcid.org/0000-0003-2534-0063), Benedikt
Heinrichs¹[\[https://orcid.org/0000-0003-3309-5985\]](https://orcid.org/0000-0003-3309-5985), and Ilona Lang¹[\[https://orcid.org/0000-0002-7202-5982\]](https://orcid.org/0000-0002-7202-5982)

¹RWTH Aachen University, Germany

²Karlsruher Institut für Technologie, Germany

Abstract:

Research institutions have established a variety of research data infrastructures that orient towards discipline or methodology specific needs of their respective research community. Technically, these infrastructures ultimately are based on off-the-shelf hardware and software building blocks – both commercial and open-source. While such enterprise ready infrastructures can scale well, they apt to data silos and typically do not adhere to scientific standards like the FAIR (Findability, Accessibility, Interoperability, Reusability) principles. Using common architecture concepts such as the FAIR Digital Object (FAIR DO) allows interconnection of these silos by adding a long term interoperability layer on top of the existing infrastructure components.

A FAIR DO is a digital representation of data as a sequence of bits, identified by a globally-unique, persistent, and resolvable identifier, described by an information record, and classified by a type. This interoperability layer on top of the existing infrastructure components allows distributed data resources to be connected and related, regardless of where they are stored.

For this purpose, we propose a shared service architecture that allows the implementation of discipline specific services and applications. The data services enable storing and processing of the data, complemented by the metadata services that describe this stored data. While separate aspects of a digital object may be stored in specialized systems like databases or object storages, a data representation layer ties these aspects together as an entity, i.e. the FAIR Digital Object. The entirety of these digital objects with their interconnections can then be represented and explored as a knowledge graph that supports specific workflows for instance machine learning or quality assurance. Finally, researchers can build on top of these workflows to realize applications relevant for their individual discipline and develop interfaces, such as connected Electronic Lab Notebooks (ELNs). These technical infrastructure need to be framed with discipline specific support, teaching and consulting activities, data governance services and harmonized agreements on interfaces, data, and metadata formats. Figure 1 gives a schematic overview of such an architecture incorporating an infrastructure and software developer's perspective.

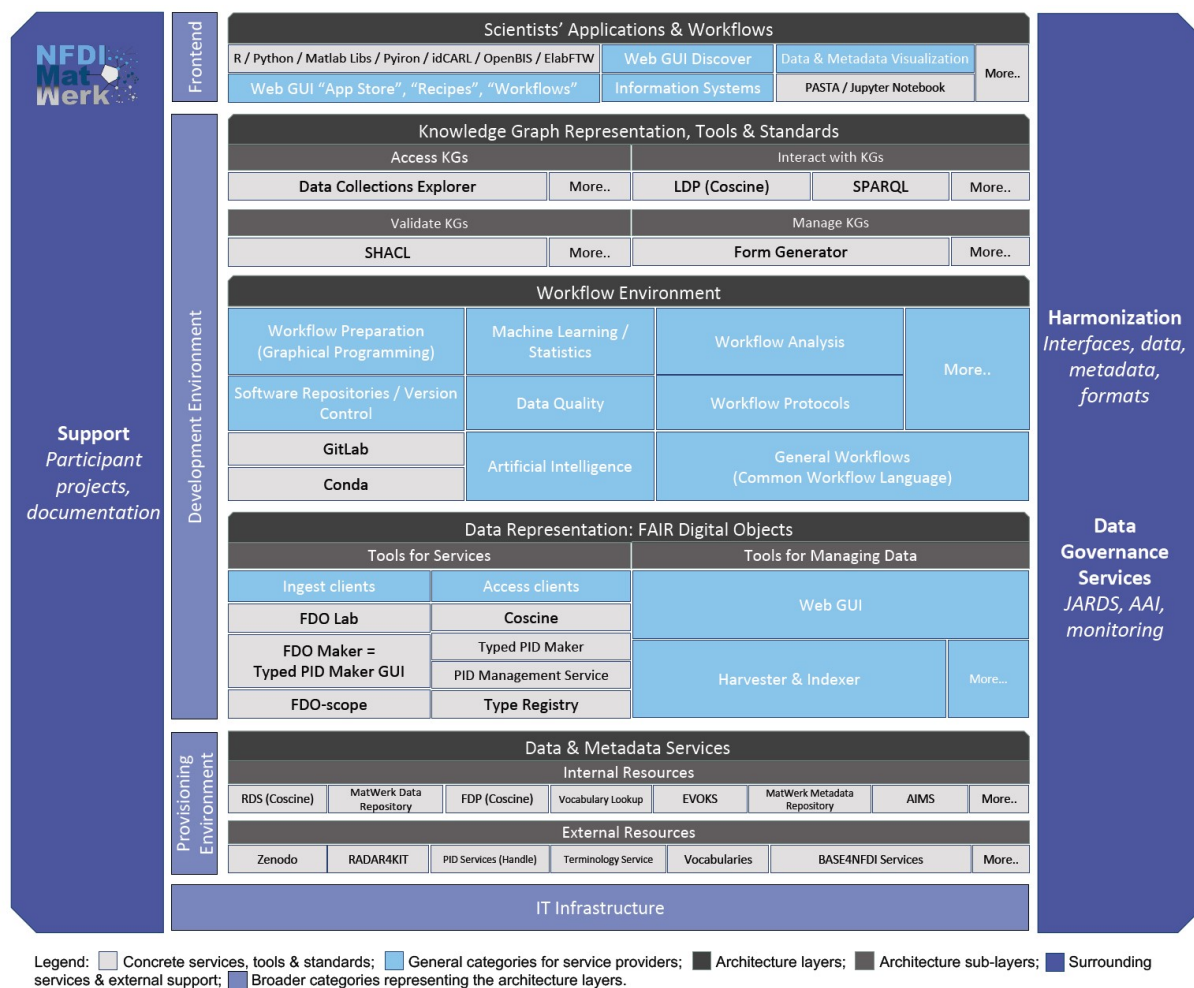


Figure 1. Schematic Overview of the envisioned architecture and its software components.

Based on a requirements analysis and researchers' needs within the NFDI-MatWerk consortium, we present a set of collectively working examples (integrated platforms: Coscine and MatWerk Data Repository, creation: FDO Maker, exploration: FAIR-DOscope) of this approach based on the concept of FAIR DO as recommended by the RDA and the European Commission. A combination of service offers and installable applications allows implementing the FAIR principles within the existing research data infrastructures. The presented results are widely based on W3C standards and are currently being evaluated in the context of the NFDI-MatWerk consortium. After that, the results will be transferred to other NFDI consortia, including NFDI4Ing, NFDI4Chem, and NFDI4Microbiota.

The approach provides a practical solution for interconnecting distributed research data infrastructures to national (like NFDI) and international (like EOSC and Gaia-X) infrastructures and preventing the creation of data silos. By allowing existing data infrastructures to make data FAIR, we enable researchers to access and reuse data from different domains, facilitating cross-disciplinary research and advancing new methods for scientific discoveries.

Contributions

- Marius Politze — Project administration, Supervision, Writing – original draft
- Yusra Shakeel — Investigation, Visualization, Software, Writing – review & editing
- Sirieam Hunke — Investigation, Visualization, Software, Writing – review & editing
- Philipp Ost — Conceptualization, Writing – review & editing
- Rossella Aversa — Project administration, Supervision, Writing – review & editing
- Benedikt Heinrichs — Software, Writing – review & editing
- Ilona Lang — Conceptualization, Investigation

The work was supported with resources granted by NFDI-MatWerk, funded by Deutsche Forschungsgemeinschaft (DFG) with project number 460247524.

Keywords: FAIR, FAIR Digital Object, Service Architecture, Standardization

The Federal State Initiatives for RDM as intermediaries in a dynamic landscape of RDM infrastructures and services

Ortrun Brand¹[\[https://orcid.org/0000-0002-6850-5123\]](https://orcid.org/0000-0002-6850-5123), Karen Bruhn²[\[https://orcid.org/0000-0002-7653-4063\]](https://orcid.org/0000-0002-7653-4063),
Magdalene Cyra³[\[https://orcid.org/0000-0001-7738-2703\]](https://orcid.org/0000-0001-7738-2703), Matthias Fingerhuth³[\[https://orcid.org/0000-0002-0248-8914\]](https://orcid.org/0000-0002-0248-8914),
Roman Gerlach⁴[\[https://orcid.org/0000-0001-5104-4247\]](https://orcid.org/0000-0001-5104-4247), Boris Jacob⁵[\[https://orcid.org/0000-0002-8565-3312\]](https://orcid.org/0000-0002-8565-3312), Cora
Krömer⁶[\[https://orcid.org/0000-0001-8473-1481\]](https://orcid.org/0000-0001-8473-1481), Ralph Müller-Pfefferkorn⁶[\[https://orcid.org/0000-0001-8719-5741\]](https://orcid.org/0000-0001-8719-5741), Heike
Neuroth⁵[\[https://orcid.org/0000-0002-3637-3154\]](https://orcid.org/0000-0002-3637-3154), Thilo Paul-Stüve²[\[https://orcid.org/0000-0001-7451-0976\]](https://orcid.org/0000-0001-7451-0976), Stephanie
Rehwald³[\[https://orcid.org/0000-0002-5884-4471\]](https://orcid.org/0000-0002-5884-4471), Janine Straka⁵[\[https://orcid.org/0000-0002-0695-1689\]](https://orcid.org/0000-0002-0695-1689), Barbara
Weiner⁷[\[https://orcid.org/0000-0003-2747-8648\]](https://orcid.org/0000-0003-2747-8648)

¹ HeFDI – Hessian Research Data Infrastructures

² FDM-SH

³ State Initiative for Research Data Management – fdm.nrw

⁴ Thuringian Competence Network for Research Data Management (TKFDM)

⁵ FRResearch Data Management in Brandenburg (FDM-BB)

⁶ bwFDM

⁷ SaxFDM – Research Data Management in Saxony

Abstract. A number of German federal states have established initiatives to support the institutionalization of RDM infrastructures and services. They can serve as intermediaries between disciplinary approaches to RDM like NFDI and RDM services in individual research institutions. This presentation gives an overview of the current state of these initiatives and provides an outlook on their role in creating synergies in RDM with a focus on their integrating potential for NFDI.

Keywords: State Initiative for RDM, Connecting RDM

Research institutions have taken up institutional approaches to providing RDM infrastructures and services at individual paces over the past years. NFDI provides a concerted disciplinary effort that despite substantial achievements has yet to integrate the wider landscape of individual researchers and institutions and to scale up infrastructures and services. Such wider adoption of standards and practices is crucial for NFDI's success.

State initiatives for RDM complement local and disciplinary approaches to RDM. As of today, the states of Baden-Württemberg, Brandenburg, Hesse, North Rhine-Westphalia, Saxony, Schleswig-Holstein and Thuringia are actively supporting institutions or networks that address RDM specifically. In some states, the state initiatives have been active from as early as 2016, preceding the inception of NFDI. Additionally, grassroots initiatives and networks exist in most of the remaining nine German states. As a whole, they form an extensive network that connects RDM-professionals and transcends disciplinary and institutional boundaries. In keeping with German federalist traditions and due to the distinctly different structural conditions in the states, the initiatives have very heterogeneous tasks and architectures [1]. They include both centralized and distributed forms of organization, and there is no uniform service portfolio.

However, they provide a broad range of infrastructures and services. Despite all differences, they share an interconnecting function that transcends individual institutions or disciplinary boundaries. This in itself makes them a valuable asset for structures like NFDI in their pursuit of engaging researchers and institutions broadly. The state initiatives have e.g. worked closely with universities of applied sciences that to date have only sparsely participated in NFDI. Through these established ties, they may markedly support efforts for their integration. Their role as intermediaries to state wide infrastructures and to the states' pertinent ministries further underscores their potential role in the coordinated development of RDM infrastructures throughout Germany [2].

In our presentation, we move from an introduction of the state initiatives to their role as intermediaries and facilitators in the RDM landscape. In doing so, we focus on their potential value for NFDI. We provide an overview of the current status and foreseeable development of the initiatives as well as the infrastructures and services for RDM that are currently offered and developed in the federal states. These include e.g. storage infrastructures and research data repositories but also consulting services and training. We suggest that the state initiatives and their affiliated services can substantially contribute to the provision and dissemination of NFDI services and outline how the state initiatives' networks and infrastructures could serve as interfaces for the integration of the broader landscape of research institutions with NFDI.

Data availability statement

The contribution does not draw on any data.

Author contributions

Project Administration: Magdalene Cyra, Matthias Fingerhuth

Conceptualization: Ortrun Brand, Karen Bruhn, Magdalene Cyra, Matthias Fingerhuth, Roman Gerlach, Boris Jacob, Cora Krömer, Ralph Müller-Pfefferkorn, Heike Neuroth, Thilo Paul-Stüve, Stephanie Rehwald, Janine Straka, Barbara Weiner

Writing Original Draft: Ortrun Brand, Karen Bruhn, Magdalene Cyra, Matthias Fingerhuth, Roman Gerlach, Boris Jacob, Cora Krömer, Ralph Müller-Pfefferkorn, Heike Neuroth, Thilo Paul-Stüve, Stephanie Rehwald, Janine Straka, Barbara Weiner

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

We thank Valerie Boda and Livia Gertis for their support in drafting the abstract for this contribution.

References

1. Alexandra Axtmann, Elisabeth Böker, Ortrun Brand, Magdalene Cyra, Nina Dworschak, Matthias Fingerhuth, Roman Gerlach, Niklas Hartmann, Gerald Jagusch, Esther Krähwinkel, Robert Lipp, Ralph Müller-Pfefferkorn, Thomas Nauss, Heike Neuroth, Markus Putnings, Stephanie Rehwald, Jessica Rex, Jürgen Rohrwild, Benjamin Slowig, Markus Putnings, Stephanie Rehwald, Jessica Rex, Jürgen Rohrwild, Benjamin Slowig, Stephan Thiemann, Barbara Weiner (2021). "Wir bringen

- die breite Basis mit" – Gemeinsames Plädoyer für eine enge Einbindung der Landesinitiativen für Forschungsdatenmanagement in die Nationale Forschungsdateninfrastruktur. Zenodo. <https://doi.org/10.5281/zenodo.4524655>
2. Sprecher:innenkreis der FDM-Landesinitiativen. (2022, October 4). FDM-Landesinitiativen und regionale Netzwerke. Vortrag bei der NFDI-Konsortialversammlung am 01. Juli 2022. Zenodo. <https://doi.org/10.5281/zenodo.6965976>

Data Trustees – They Do Work! The Example of Research Data Centers

Daniel Fuß¹ and Marie-Christine Laible²

¹ Leibniz Institute for Educational Trajectories, Germany (Research Data Center)

² Federal Office for Migration and Refugees (Research Data Center)

Abstract. This contribution presents the long established system of accredited Research Data Centers (RDCs). Created in the data-landscape of the social, behavioral, educational, and economic sciences, they enable access to restricted data and bridge interests of data providers and researchers. A distinctive feature of most research data in the above disciplines is the coverage of real persons. Such sensitive data require specific safeguards. The focus of this contribution is on the institutionalized processes of connecting and securing appropriate research data management strategies for this sort of data. It includes quality assurance measures through the accreditation of RDCs, a monitoring system and the regular cooperation of accredited RDCs.

Keywords: Research Data Infrastructure, Quality Assurance, Sensitive Data, Forschungsdatenzentrum, RatSWD, KonsortSWD

1. Overview

An increasingly prominent term in the discussion about the provision of sensitive data for research purposes is that of the ‘Data Trustee’ as data intermediation services (cf. EU Data Governance Act [1]). Data trustees mediate the interests of data providers or data producers and data users. They should ensure a balance between the high level of protection for the often personal data on the one hand and the information depth necessary for high-quality research on the other hand by means of pseudonomization or anonymization processes, contract management etc. These demands largely describe the range of tasks of accredited RDCs. Thus, they are a model of data trustees. In our presentation we take an exemplary look at the rather young RDC of the Federal Office for Migration and Refugees (BAMF-FDZ), which prepares and makes available data from the Central Register of Foreigners (Ausländerzentralregister, AZR) for research purposes.

The network of decentralized RDCs is unique in Germany. RDCs make up an infrastructure that successfully accommodates the different demands of researchers and that has continuously adjusted to new requirements. They are considered to be “best practice” by the Council for Scientific Information Infrastructures (RfII [2, pp. 30–33], cf. also Wissenschaftsrat [3, pp. 81–82]) and have played a growing role in the German scientific system during the last two decades. The first RDC was founded in 2001 by the Federal Statistical Office (FDZ-Bund), followed by five more RDCs until 2008 – the year in which the German Data Forum (RatSWD) introduced the accreditation process. Currently, there are 41 accredited RDCs with a thematic range from economic and insurance data to education and labor market data to health and social data as well as a broad spectrum of data types from large-scale panel studies such as the Socio-Economic Panel Study (SOEP) and the National Educational Panel Study (NEPS) to register and census data to rich qualitative data collections in text, audio, and video formats.

In 2021, these RDCs had 5,517 datasets made available to nearly 53,000 national and international data users, counted more than 88,000 downloads of open access datasets, and recorded 3,101 scientific publications based on RDC data (cf. RatSWD [4]).

All accredited RDCs are members of the Committee for Data Access (CDA). It is made up of the heads of the RDCs and convenes twice a year to share best practices in research data management, but also to exchange experiences with innovations. All efforts of the CDA aim at safeguarding the continuous improvement of the research data infrastructure, which includes advancing the quality and quantity of available data as well as facilitating data access for researchers. Specific working groups are set up to deal with cross-RDC issues and to develop concerted answers to common (user) demands. A coordination office acts as single point of consultation and contact. The RDCs in the CDA are also essential in the context of the Consortium for the Social, Behavioral, Educational and Economic Sciences (KonsortSWD) within the National Research Data Infrastructure (NFDI), particularly due to their central role in various measures around research data management issues.

The CDA cooperates closely with the RatSWD. First, the RatSWD accredits new RDCs based on recommendations from the CDA. Accreditation is bound to some commonly defined mandatory criteria and a couple of standardized information requests. Thus, RDC accreditation ascertains a minimum level of data management and data sharing based on common guidelines (Bug et al. [5]). RatSWD and CDA have also established a reporting and monitoring system. All accredited RDCs submit an annual questionnaire reviewing the use of their data, their services and activities in the previous year. Finally, a general complaints office has been installed for data users to report alleged infringements of accreditation criteria. All three measures are important quality assurance instruments. By accrediting and monitoring RDCs, the CDA and the RatSWD have created a highly connected and sustainable research data infrastructure that guarantees user-friendly research data management and secure data access.

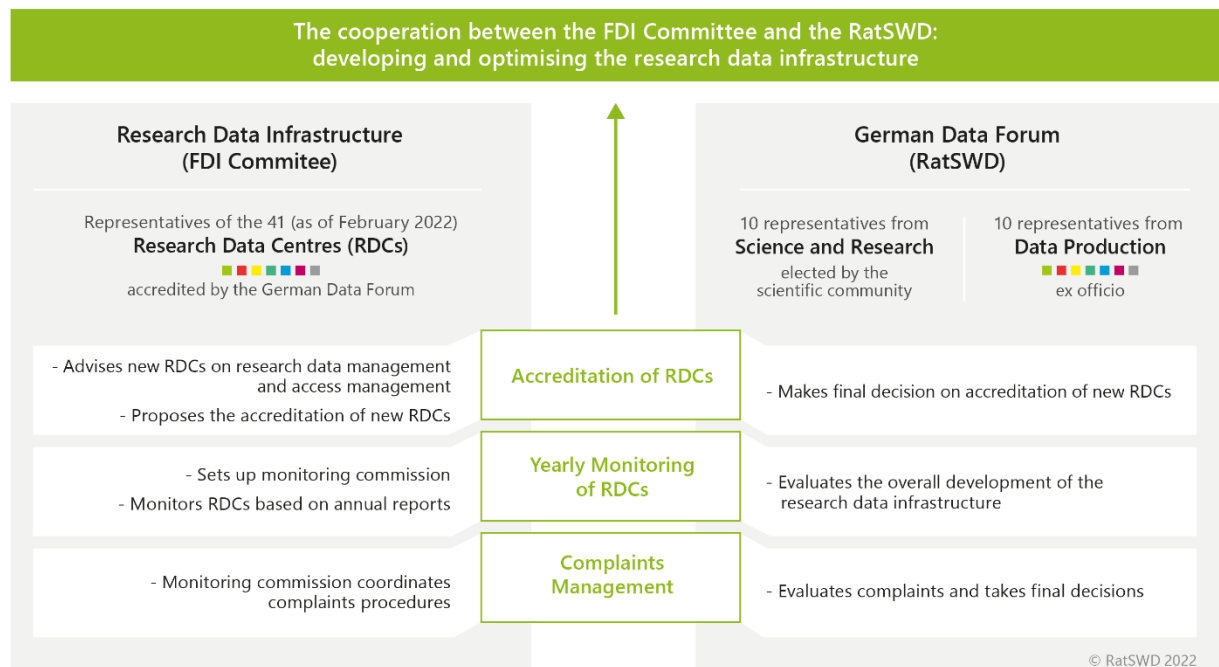


Figure 1. Connections of CDA and RatSWD for quality assurance of research data management in the social, behavioral, educational, and economic sciences.

Data availability statement

-

Underlying and related material

-

Author contributions

-

Competing interests

The authors declare that they have no competing interests.

Funding

-

Acknowledgement

-

References

1. European Union. "Regulation (EU) 2022/868 of 30 May 2022 on European data governance, and amending Regulation (EU) 2018/1724 (Data Governance Act)," Official Journal of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868> (16 April 2023)
2. RfII, 2016. „Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland.“ Göttingen, Rat für Informationsinfrastrukturen. <http://www.rfii.de/?p=1998> (16 April 2023)
3. Wissenschaftsrat, 2011. „Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften.“ Berlin. https://www.wissenschaftsrat.de/download/archiv/10465-11.pdf;jsessionid=947594311989891E2366B5CB396E9580.delivery2-master?_blob=publicationFile&v=3 (16 April 2023)
4. RatSWD, 2022. „Research Data Infrastructure accredited by the RatSWD.“ Berlin, German Data Forum. <https://doi.org/10.17620/02671.71> (16 April 2023)
5. M. Bug, S. Liebig, C. Oellers and R. T. Riphahn, „Operative und strategische Elemente einer leistungsfähigen Forschungsdateninfrastruktur in den Sozial- und Wirtschaftswissenschaften,“ Jahrbücher für Nationalökonomie und Statistik, 238, 6, pp. 571–590, 2018, doi: <https://doi.org/10.1515/jbnst-2018-0029>

Extended Abstract

Coscine.nrw Landesweite Basisversorgung zur Verwaltung von Forschungsdaten im Open Source Modell

Marius Politze¹[\[https://orcid.org/0000-0003-3175-0659\]](https://orcid.org/0000-0003-3175-0659), Ilona Lang¹[\[https://orcid.org/0000-0002-7202-5982\]](https://orcid.org/0000-0002-7202-5982), and
Katja Jansen¹[\[https://orcid.org/0009-0005-7076-9848\]](https://orcid.org/0009-0005-7076-9848)

¹RWTH Aachen University, Germany

Abstract: n.a.

Keywords: FAIR, FAIR Digital Object, Data Storage, Metadaten, Scrum

1 Abstract

An der RWTH Aachen wird seit 2018 die Forschungsdatenplattform Coscine als Open-Source-Software entwickelt und für die Verwaltung von Forschungs(meta)daten, sowie zur Kontingentierung und Provisionierung von Speicherressourcen für Forschungsdaten eingesetzt. Coscine ist dazu gemäß der FAIR-Prinzipien entwickelt und implementiert Schnittstellen für sog. FAIR Digital Objects. Für Forschende bietet Coscine den Zugriff auf alle Forschungsdaten eines Forschungsprojekts, die Verknüpfung mit projekt- oder fachspezifischen Metadaten sowie die Verwaltung von Projektmitgliedern. Dank des niederschweligen Zugangsmanagements kann Coscine als Kollaborationsplattform über Hochschulgrenzen hinaus verwendet werden. In die Entwicklung von Coscine fließen zudem die Erkenntnisse und Anforderungen aus nationalen (NFDI, NHR) und internationalen Vorhaben (EOSC, gaia-x, RDA) ein und ermöglichen sowohl eine fachspezifische als auch eine fachübergreifende Verwendung der Plattform. Im Rahmen des Serviceangebots Coscine.nrw wird die Software Coscine für alle Hochschulen der DH.NRW zur Verfügung gestellt.

1.1 Funktionsumfang der Datenmanagementplattform

Coscine bietet Forschenden eine Vielzahl von Funktionalitäten zur Verwaltung der Forschungsdaten unter Berücksichtigung der FAIR-Prinzipien. Coscine bietet dazu eine Abbildung einer Projektstruktur zur Verwaltung von Zugriffsberechtigungen und der Forschungs(meta)daten. Durch die Integration verschiedener Services können Forschende alle Projektdaten an einem Ort sehen und verwalten (vgl. Figure 1).

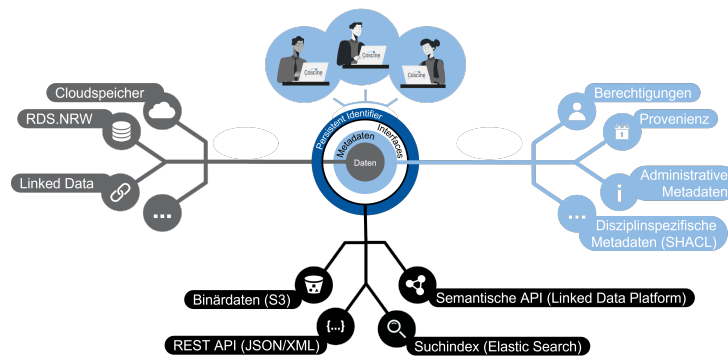


Figure 1. Auf Basis von definierten Schnittstellen kombiniert Coscine Speicherressourcen und Metadaten und stellt auf dieser Basis die Umsetzung der FAIR-Prinzipien nach dem Konzept des FAIR-Digital Object sicher.

1.1.1 Authentifizierung

Coscine nutzt die Authentifizierungs- und Autorisierungsinfrastruktur des DFN (DFN-AAI). Die DFN-AAI ist eine Kernkomponente der IAM4NFDI, sodass sich Coscine.nrw in die bestehende Landschaft integriert. Zur Authentifizierung von Nutzenden, die nicht direkt mit einer Hochschule oder Forschungseinrichtung in der DFN-AAI assoziiert sind, bietet Coscine die Möglichkeit der Authentifizierung über die Open Researcher and Contributor ID (ORCID).

1.1.2 Speicherplatzverwaltung

Der Zugang zu Speicherplatz auf dem Landesweiten Forschungsdatenspeicher RDS.nrw wird über ein wissenschaftsgeleitetes Antragsverfahren vergeben. Zur Sicherstellung der Einhaltung der guten wissenschaftlichen Praxis (GWP) und der Umsetzung der FAIR-Prinzipien wird das jeweilige FDM-Konzept im Rahmen eines Datenmanagementplans dargelegt.

Neben RDS.nrw können auch GitLab (für textbasierte Forschungsdaten bspw. Softwarequellcodes) oder Linked Data (für eine Metadatenverwaltung und Referenzierung externer Daten) Ressourcen erzeugt werden. Weitere (Community-) Cloud-Anwendungen, wie Sciebo und Nextcloud, sind geplant (vgl. Figure 1).

1.1.3 Metadatenverwaltung

Das Metadatenmanagement auf Projekt-, Ressourcen- und Dateiebene basiert auf den W3C Standards RDF und der Shapes Constraint Language (SHACL) und sichert so die Interpretierbarkeit und Interoperabilität der Metadaten. Alle von Coscine verwalteten Elemente erhalten einen Persistenten Identifier (PID). Nutzende können aus fach- und projektspezifischen Metadatenprofile ausgewählt. Die individuelle Erstellung von Metadatenprofilen ist dank der Einbindung des AIMS Projekts auch ohne fundierte Kenntnisse im Bereich RDF und SHACL möglich. Coscine bietet so die Möglichkeit, die in verschiedenen Fachbereichen und -Konsortien erstellten Ontologien wiederzuverwenden oder zu kombinieren.

Als Anknüpfungspunkt für internationale Infrastrukturen implementiert Coscine FAIR Data Point (FDP). Diese ist zu den FAIR Digital Objects kompatibel und basiert technisch auf dem Resource Description Framework (RDF), dem Data Catalog Vocabulary (DCAT) und der Linked Data Platform (LDP).

1.1.4 Automatisierung

Coscine bietet Schnittstellen zur Automatisierung von Arbeitsprozessen. Über die REST-API können Dateien beispielsweise automatisch mit Metadaten versehen und hochgeladen werden. Durch die Möglichkeit, über S3-Clients direkt mit verwalteten Objektdatenspeichern zu interagieren, wird eine weitere Automatisierungsmöglichkeit angeboten.

1.1.5 Archivierung

Neben der Speicherung von sogenannten aktiv genutzten heißen Forschungsdaten ermöglicht Coscine in Kombination mit RDS.nrw die Archivierung von kalten Forschungsdaten und Metadaten und wird für 10 Jahre nach Projektende gemäß der GWP gewährleistet. Danach sind der Übergang in ein geeignetes Langzeitarchivierungssystem und Maßnahmen für die Langzeitverfügbarkeit notwendig.

1.2 Communitygetriebene Weiterentwicklung im Open Source Modell

Die Entwicklung von Coscine unter einer freien Lizenz im Rahmen eines öffentlichen GitLab Projekts ermöglicht allen interessierten Nutzenden eine Teilhabe und Einbringung in die Coscine-Entwicklung.

Die Umsetzung neuer strategischer Anforderungen wird dabei über Prozesse aus dem Scaled Agile Framework (SAFe) und der agilen Scrum Methode an der RWTH Aachen gemanagt (vgl. Figure 2). Neue Anforderungen werden als sogenannte Epics formalisiert, in denen die benötigten Weiterentwicklungen im Rahmen einer vorgefertigten Vorlage ausgearbeitet werden. Die Epics sind öffentlich einsehbar und werden in enger Kommunikation mit allen beteiligten Stakeholdern auf Richtigkeit, Vollständigkeit, Mehrwert für Nutzenden sowie zeitlicher Kritikalität geprüft. Eine interne Steuerungsgruppe prüft in einem zweiten Schritt die Umsetzbarkeit und strategische Einordnung der Epics. Die nachfolgende technische Analyse des Epics erfolgt durch das Scrum-Team. Anschließend werden alle neuen Epics basierend auf der WSJF-Formel priorisiert, wobei ein Strategiefaktor, der Profit für alle Nutzenden, das Risikomanagement, die zeitliche Kritikalität und der Arbeitsaufwand berücksichtigt wird. Die drei Epics mit der höchsten Priorisierung werden an den/das Scrum-Team zur Entwicklung übergeben. Nach erfolgter Entwicklung wird die Umsetzung durch das Scrum-Team den Stakeholdern präsentiert und an alle Nutzenden kommuniziert.

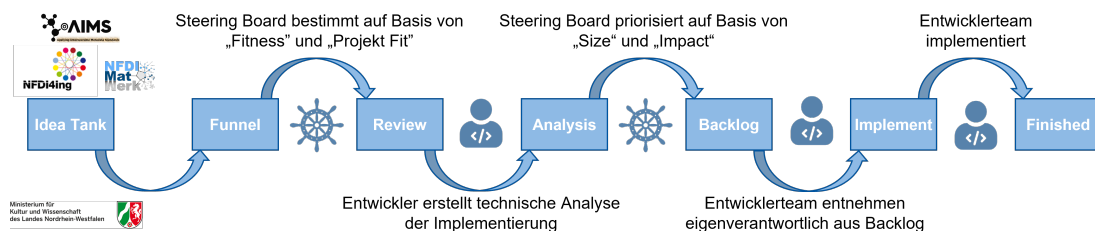


Figure 2. Input von den verschiedenen Stakeholdern wird im SAFe Prozess überprüft, priorisiert und umgesetzt.

Ein zusätzlich eingerichtetes Contributions-Repository bietet der Community eine Plattform für das Teilen eigener Entwicklungen und Anwendungen, die beispielsweise im Rahmen der Nutzung der offenen Schnittstellen von Data Stewards auf Projektbene beigetragen werden. Neben dem Teilen von Code ermöglicht die Plattform auch

einen Austausch für Forschende und Data Stewards an den verschiedenen Standorten zur gegenseitigen Unterstützung für die Automatisierung von individuellen Prozessen.

1.3 Zusammenfassung und Ausblick

Dank offener Schnittstellen (REST-APIs), der fortwährenden Anbindung an anerkannte Konzepte wie FAIR Digital Objects ist Coscine ein wichtiger Baustein, um verteilte Speicherinfrastrukturen entsprechend der FAIR Prinzipien weiterzuentwickeln. Durch die Metadaten-Verwaltung von Coscine können Speichersysteme, die Industriestandard entsprechen, gemäß der FAIR-Prinzipien verwendet werden. Dies verbessert insbesondere die Partizipationsmöglichkeiten kleinerer Hochschulen an diesen wissenschaftlichen Infrastrukturen und erhöht so langfristig die Wirtschaftlichkeit der investierten Ressourcen. Die communitygetriebene Weiterentwicklung ermöglicht die Beteiligung einer Vielzahl von Stakeholdern an der Entwicklungsroadmap.

Die Software Coscine wird in NRW als Serviceangebot Coscine.nrw etabliert und steht mit dem beschriebenen Funktionsumfang den Hochschulen in NRW zur Verfügung. Darüber hinaus können auch Use Cases der NFDI-Konsortien außerhalb von NRW Coscine als Basisinfrastruktur nutzen, um Forschungsdaten zu verwalten oder Forschungsdatenspeicher zu kontingentieren und provisionieren.

Beiträge der Autoren

Marius Politze — Project administration, Supervision, Writing – original draft
Ilona Lang — Conceptualization, Investigation, Writing – original draft
Katja Jansen — Writing – review & editing

Förderung

Die Arbeiten wurden mit Ressourcen von NFDI4Ing, gefördert von der Deutschen Forschungsgemeinschaft (DFG) unter Projektnummer 442146713, NFDI-MatWerk, gefördert von der Deutschen Forschungsgemeinschaft (DFG) unter Projektnummer 460247524 unterstützt. Coscine.nrw ist ein Service des Ministeriums für Kultur und Wissenschaft des Landes Nordrhein-Westfalen.

NFDI4Health Local Data Hubs for Finding and Accessing Health Data

Making Distributed Data Accessible through a SEEK-Based Platform

Frank Meineke¹[\[https://orcid.org/0000-0002-9256-7543\]^{*}, Martin Golebiewski²\[\\[https://orcid.org/0000-0002-8683-7084\\]^{*}, Xiaoming Hu²\\[\\\[https://orcid.org/0000-0002-8318-3222\\\]\\]\\(https://orcid.org/0000-0002-8318-3222\\), Toralf Kirsten^{1,3}\\[\\\[https://orcid.org/0000-0001-7117-4268\\\]\\]\\(https://orcid.org/0000-0001-7117-4268\\), Matthias Löbe^{1,3}\\[\\\[https://orcid.org/0000-0002-2344-0426\\\]\\]\\(https://orcid.org/0000-0002-2344-0426\\), Sebastian Klammt^{1,3}\\[\\\[https://orcid.org/0000-0001-7852-4769\\\]\\]\\(https://orcid.org/0000-0001-7852-4769\\), Ulrich Sax⁴\\[\\\[https://orcid.org/0000-0002-8188-3495\\\]\\]\\(https://orcid.org/0000-0002-8188-3495\\), Wolfgang Müller²\\[\\\[https://orcid.org/0000-0002-4980-3512\\\]\\]\\(https://orcid.org/0000-0002-4980-3512\\) on behalf of the NFDI4Health consortium\]\(https://orcid.org/0000-0002-8683-7084\)](https://orcid.org/0000-0002-9256-7543)

¹ Leipzig University, Germany

² Heidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, Germany

³ Leipzig University Medical Center, Germany

⁴ University Medical Center Göttingen, Germany

^{*}contributed equally

Abstract: To support federated data structuring and sharing for sensitive health data from clinical trial, epidemiological and public health studies in the context of the German National Research Data Infrastructure for Personal Health Data (NFDI4Health), we have developed Local Data Hubs (LDHs) based on the FAIRDOM-SEEK platform. Those LDHs connect to the German Central Health Study Hub (CSH) to make the health data searchable and findable. This decentralised approach supports researchers to make health studies with their data FAIR (Findable, Accessible, Interoperable and Reusable), and at the same time fully preserves data protection for sensitive data.

Keywords: Life Sciences, Local Data Hub, FAIRDOM, SEEK, Enabling RDM, Data sharing

1. Background and motivation

NFDI4Health is a consortium of the German National Research Data Infrastructure (NFDI) Initiative. The aim is to establish an overarching infrastructure for person-related health study data in Germany [1]. The consortium is developing tools and services to make data and metadata from clinical trials, epidemiological and public health studies publicly available to share them with the scientific community.

NFDI4Health takes a distributed and federated approach: While the primary search platform for health studies and corresponding metadata, the German Central Health Study Hub (CSH), is provided centrally - including detailed, fine-grained search capabilities - the actual data remain decentralised at the Data Holding Organisations (DHO) under the full control of the data owners. To support this overarching query and make study data findable and accessible, it is necessary to harmonise the corresponding metadata so that they follow the same formatting and structure, as well as use the same terminologies, i.e. to make data interoperable with each other. This is done through the development and use of a common metadata schema (MDS) [2] tailored to the needs of and developed by NFDI4Health. This MDS is implemented

in the central and decentralised services of NFDI4Health to ensure interoperability of those infrastructure components.

This paper focuses on the development the local research data management service for decentralised data sharing in NFDI4Health, the so-called Local Data Hubs (LDH), and the adaptation of its underlying SEEK-based platform.

2. Health research data management via federated local data hubs

The use of data hubs for clinical and epidemiological trials and data is motivated by the Leipzig Health Atlas platform [3,4], which is based on the widely used FAIRDOM-SEEK (SEEK) system [5,6].

2.1 The FAIRDOM-SEEK platform for scientific data management

SEEK is a data sharing platform that is developed by the FAIRDOM community [7] and allows storing and registration of research data and corresponding information, as well as sharing the corresponding metadata in a structured fashion [8]. SEEK is an open source web-based data management and commons platform, for sharing scientific research datasets, models or simulations, processes and research outcomes. It preserves associations between them, along with information about the people and organisations, as well as other information about the related research projects.

SEEK provides a detailed approach to access control: Users can keep their uploaded documents and data completely private, share them between individuals or across projects, or make them available to the public.

2.2 The NFDI4Health Local Data Hub (LDH)

For the Local Data Hubs (LDH) (see figure 1) we have implemented the NFDI4Health metadata schema (MDS) in SEEK for metadata descriptions of studies and data collections, as well as their corresponding data sets and documents such as standard operating procedures (SOPs), publications, trial protocols, instruments, etc. (see figure 1). An LDH also allows registration of events and presentations. Content may be grouped by programme and project layer, as well as by investigations, studies and resources (based on the implementation of the ISA metadata tracking framework [9] in SEEK) and linked to the researchers who created it. We offer a set of best practice guidelines for researchers who want to make their data available and usable to the widest possible audience. Spreadsheet and imaging data, as well as other commonly used formats even allow directly viewing in the browser without download, with additional advanced features for spreadsheets.

2.3 The NFDI4Health LDH development process

To make the best use of the LDH in the context of NFDI4Health, SEEK is being optimised and extended to fulfil the specific requirements for health data. Wherever possible, we aim to integrate extensions into the main SEEK development branch. An example of this is the extension to allow users to define additional metadata elements and attributes as "custom metadata" which is used to add NFDI4Health-specific metadata to the general metadata defined by the SEEK system. The programmatic interface (API) of SEEK was also adapted to support the transfer of such extended metadata.

This new feature is used for metadata export to the central component of the NFDI4Health infrastructure (CSH). Currently this export builds on the SEEK JSON API structure, but will be possible as HL7 FHIR® transfer [10] as soon as a FHIR exporter module in

SEEK becomes available that currently is built on the profiling of the NFDI4Health MDS for SEEK [11].

3. Conclusions and Future Work

Currently, several instances of the Local Data Hub are online: A public test platform for the evaluation of new functions, the productive instance of the LDH Leipzig with the introduced name "Health Atlas" and a sample of data from the Clinical Trial Centre (ZKS) Leipzig. They differ in content, public access and level of implementation of the MDS. It is planned to have 6 LDHs in the current year, mostly associated with central supporting structures for clinical trials at university medical centres in Germany.

This work demonstrates that FAIRDOM-SEEK is well suited to support research data management and sharing in general, and also applicable for health data, despite the system's original focus on experiments in systems biology and 'omics' data. With the adaptation to specific requirements for metadata from health studies we have demonstrated that SEEK is flexible enough to be used also in other domains. The system might be extended not only to the health domain, but also for non-clinical / non-medical data, e.g. in the context of NFDI for FAIR research data management.

Nevertheless, different communities and different scenarios may require different LDH flavours including some community specific best practice infrastructure elements.

The screenshot displays the Leipzig Health Atlas (LHA) interface. At the top, there is a navigation bar with the NFDI4Health and Leipzig Health Atlas logos, a search bar, and links for 'About', 'Help', 'Register', and 'Log in'. Below the navigation bar, there are tabs for 'Details' and 'Related items'. The main content area features a large heading for the study: '#1 Characterization of multimorbidity patterns and association with health outcomes in the elderly'. Underneath, there is a 'Motivation' section, a list of data sources (UNIGE, SAS, UCSC, UP, IACS), and inclusion criteria. A sidebar on the right shows 'Creators and Submitter' (Matthias Löbe) and 'Activity' (Views: 818). The footer contains the 'Health Atlas - Local Data Hub/Leipzig' logo and contact information.

Figure 1. The Leipzig Health Atlas (LHA) as an instance of the LDH. The screenshot shows an example for a study in the LHA with its study resources (e.g. clinical datasets).

Data availability statement

SEEK and the NFDI4Health LDH source code is open source and can be found at GitHub:
 FAIRDOME-SEEK source code: <https://github.com/seek4science/seek>
 NFDI4Health LDH source code: <https://github.com/nfdi4health/ldh>

Author contributions

All authors collaborate in the NFDI4health project. FM and MG prepared the manuscript, all authors reviewed and finalized.

Competing interests

The authors declare that they have no competing interests.

Funding

We greatly appreciate the funding from the Deutsche Forschungsgemeinschaft (DFG) through projects no. 442326535 (NFDI4health), 451265285 (NFDI4health TF COVID19), and 315072261 (NMDR2), as well as from the Klaus Tschira Foundation (KTS).

Acknowledgement

We thank the whole NFDI4Health consortium for their valuable input.

References

1. Fluck, J., Lindstädt, B., Ahrens, W., Beyan, O., Buchner, B., Darms, J., Depping, R., Dierkes, J., Neuhausen, H., Müller, W., Zeeb, H., Golebiewski, M., Löffler, M., Löbe, M., Meineke, F., Klammt, S., Fröhlich, H., Hahn, H., Schulze, M., Pischon, T., Nöthlings, U., Sax, U., Kusch, H., Grabenhenrich, L., Schmidt, C.O., Waltemath, D., Semler, S., Gehrke, J., Kirsten, T., Praßer, F., Thun, S., Wieler, L., Pigeot, I., "NFDI4Health – Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten", BFDM, Nr. 2, S. 72–85, Juli 2021, doi: <https://doi.org/10.17192/bfdm.2021.2.8331>
2. A. Shutsko, C. O. Schmidt, S. A. I. Klopfenstein, J. Darms, M. Golebiewski, C. N. Vorisek, H. Abaza, NFDI4Health Task Force COVID-19, NFDI4Health. "Metadata schema of the NFDI4Health and the NFDI4Health Task Force COVID-19 (V3_0)." doi: <https://doi.org/10.4126/FRL01-006439110>
3. T. Kirsten, F.A. Meineke, H. Loeffler-Wirth, C. Beger, A. Uciteli, S. Stäubert, M. Löbe, R. Hänsel, F.G. Rauscher, J. Schuster, T. Peschel, H. Herre, J. Wagner, S. Zachariae, C. Engel, M. Scholz, E. Rahm, H. Binder, M. Loeffler, LHA team, "The Leipzig Health Atlas - An Open Platform to Present, Archive, and Share Biomedical Data, Analyses, and Models Online", *Methods Inf Med.* 61, S 02, e103-e115, December 2022, doi: <https://doi.org/10.1055/a-1914-1985>
4. "The Leipzig Health Atlas": <https://www.health-atlas.de> (accessed 21 April 2023)
5. K. Wolstencroft, S. Owen, O. Krebs, Q. Nguyen, N.J. Stanford, M. Golebiewski, A. Weidemann, M. Bittkowski, L. An, D. Shockley, J.L. Snoep, W. Mueller, C. Goble, "SEEK: a systems biology data and model management platform", *BMC Systems Biology*, 9, 33, July 2015, doi: <https://doi.org/10.1186/s12918-015-0174-y>
6. "FAIRDOM-SEEK: An open source web-based cataloguing and commons platform." <https://seek4science.org/>(accessed 21 April 2023)
7. "FAIRDOM - A Consortium of Services for Research Data Management and More": <https://fair-dom.org/>(accessed 21 April 2023)
8. K. Wolstencroft, O. Krebs, J.L. Snoep, Stanford, N.J. Bacall F, M. Golebiewski, R. Kuzyakiv, Q. Nguyen, S. Owen, S. Soiland-Reyes, J. Straszewski, D.D. Van Niekerk,

A.R. Williams, L. Malmström, B. Rinn, W. Müller, C. Goble, "FAIRDOMHub: a repository and collaboration environment for sharing systems biology research", *Nucleic Acids Research*, 45, D1, pp. D404-D407, November 2016, doi:

<https://doi.org/10.1093/nar/gkw1032>

9. "ISA Commons": <https://www.isacommons.org> (accessed 21 April 2023)
10. "HL7 FHIR (Fast Healthcare Interoperability Resource)": <https://www.hl7.org/fhir/> (accessed 25 April 2023)
11. S. A. I. Klopfenstein, C. N. Vorisek, A. Shutsko, M. Lehne, J. Sass, M. Löbe, & C.O. Schmidt, S. Thun, "Fast Healthcare Interoperability Resources (FHIR) in a FAIR Metadata Registry for COVID-19 Research", *Stud Health Technol Inform.* 287, pp. 73-77, November 2021, doi: <https://doi.org/10.3233/SHTI210817>

RDMkit: The Research Data Management Toolkit for Life Sciences

Nazeefa Fatima¹[\[https://orcid.org/0000-0001-7791-4984\]](https://orcid.org/0000-0001-7791-4984), Pinar Alper²[\[https://orcid.org/0000-0002-2224-0780\]](https://orcid.org/0000-0002-2224-0780), Federico Bianchini¹[\[https://orcid.org/0000-0002-9016-4820\]](https://orcid.org/0000-0002-9016-4820), Korbinian Bösl³[\[https://orcid.org/0000-0003-0498-4273\]](https://orcid.org/0000-0003-0498-4273), Ulrike Wittig⁴[\[https://orcid.org/0000-0002-9077-5664\]](https://orcid.org/0000-0002-9077-5664), Carole Goble⁵[\[https://orcid.org/0000-0003-1219-2137\]](https://orcid.org/0000-0003-1219-2137), Frederik Coppens⁶[\[https://orcid.org/0000-0001-6565-5145\]](https://orcid.org/0000-0001-6565-5145)

¹ University of Oslo, Norway

² Luxembourg National Data Service, Luxembourg

³ University of Bergen, Norway

⁴ Heidelberg Institute for Theoretical Studies, Germany

⁵ University of Manchester, United Kingdom

⁶ The Flemish Institute for Biotechnology, Belgium

Abstract. Effective data management extends over the entire life cycle of data, from the point of creation through to dissemination and archiving, and usually continues long after a research project is concluded. Tailored guidance for management of research data is increasing its importance among data-driven scientific investigations. There is now increasing demand for Research Data Management (RDM), by funders and host institutions, for researchers to develop and implement data management plans for projects. The RDMkit, by ELIXIR Europe, is an open community-led resource designed to share knowledge on how to carry out RDM in life sciences research. It fills the training and learning gap by providing knowledge on domain-focused RDM considerations, task-focused solutions and resources, and tool assemblies that showcase real-life examples addressing how combinations of tools can be used to go through the data life cycle. The technical infrastructure of the RDMkit is reproducible, making it possible for other organisations to use the structure as a guide and inspiration to set up RDM source for their users. In this talk, RDMkit will be introduced with example best practice guidelines, and integrations and knowledge sources.

Keywords: Crowdsourced, RDM, Knowledge Resource

1. RDMkit Overview

Research data management (RDM) is crucial for the reproducibility of studies and the re-usability of scientific results. RDM is based on a life-cycle view of data that extends beyond the timespan of a research study, preceding data's creation and often lasting beyond data and results dissemination. In order to foster RDM and embed it into the scientific process, funders require researchers to systematically plan for data management with Data Management Plans (DMPs).

Funders, research institutions and infrastructures, offer a multitude of support to help researchers in building and executing DMPs. The support landscape includes institutional data stewards who consult and train researchers, as well as RDM policies, guidelines, standards

and software tools from various sources. Two aspects of current support prevent it from reaching its full potential. First, for researchers, navigating this populous landscape, even within a single research infrastructure, can be overwhelming. Data stewards play a role here as advisors, however, they need mechanisms to effectively capture and disseminate their know-how. Secondly, most guidance found in current resources is often generic focusing on funder DMP templates, whereas surveys reveal that researchers need tailored, discipline-specific guidance and examples [1].

To address these challenges and bridge the knowledge gaps, ELIXIR (the pan-European research infrastructure for life-science data [2]) has convened a transnational community of bioscientists to build the RDMkit [3]. Launched in March 2021, the RDMkit is a collaborative effort by data stewards, life scientists, and RDM experts from all areas of biomedical sciences. It provides RDM best practice guidelines reachable through entry points based on the data lifecycle, the personas in RDM as well as various life science domains. Per topic, RDMkit highlights key considerations and the resources that can be used to build solutions in a way that makes life-science data Findable, Accessible, Interoperable and Reusable (FAIR). The RDMkit integrates with other ELIXIR tools and resources such as FAIRCookbook [4], FAIRsharing [5], bio.tools [6], Data Stewardship Wizard [7] and TeSS [8] training portal enabling the user to deliver a seamless data management plan and form an RDM knowledge commons for the ELIXIR community.

For researchers, the RDMkit is a one-stop open source of information, advice, and signposting to RDM know-how, tools, examples and best practices written by life scientists for life scientists. For data managers, RDMkit is a resource to complement institutional guidelines. For funding agencies and policymakers, RDMkit is a resource that can be included in guidelines. With over 100 webpages, RDMkit content spans a range of life science domains from metagenomics to human data and various tasks from metadata collection to data publication, considering all steps of the data management life cycle. In addition to generic advice, the RDMkit provides a collection of country-specific information resources such as local policies and national regulations on data ethics.

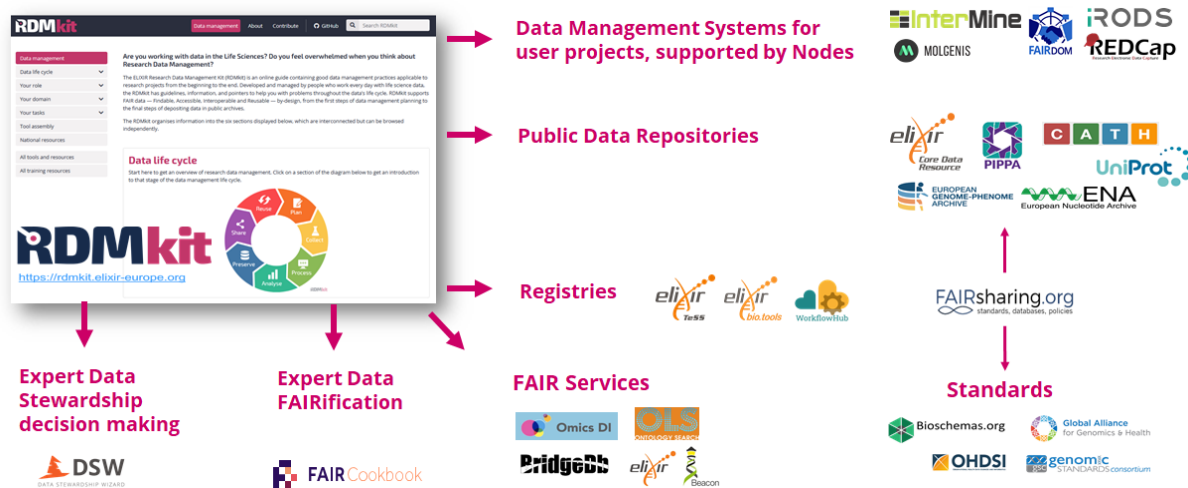


Figure 1. RDMkit offers exchange of FAIR know-how through integrations with other databases and resources, and signposting to resources in the community.

As a community-driven effort, coordinated by the Editorial Board, RDMkit is open to contributions from everyone. So far, the content is developed by over 150 contributors with references to over 340 tools and resources explained in the context of data management solutions to real-world problems. Other European Research Infrastructure providers - EuroBioImaging, BioExcel Centre of Excellence for Molecular Modelling - have contributed to specialist pages.

RDMkit has immediately gathered high visibility with 10K+ users from across 100 countries since its launch in March 2021 (Google Analytics), with users from across South America and the African continents as well as the USA and Europe. RDMkit is scalable and replicable, with open access code and content material available at GitHub. Although RDMkit pages are written in English, there is no limit to the languages for the content development.

Shortly after its release, the RDMkit has been recognised in funder guidelines in the EU. The European Commission's Horizon Europe Programme Guide recommends it as the "resource for Data Management guidelines and good practices for the Life Sciences", and it is listed in European Research Council guidelines for grantees. The RDMkit website is referred to by various national funders, universities and research institutes. Its open-source web infrastructure has been adopted by several national/international RDM initiatives, ranging from Australian BioCommons, to the 1+ Million Genomes Trust Framework, and the European BY-COVID project. The NIH Office of Data Strategy has representation on the RDMkit editorial board, for the possible alignment of NIH-funded RDM support activities with the RDMkit approach.

At the 1st Conference on Research Data Infrastructure (CoRDI), the RDMkit will be introduced with example best practice guidelines from its different sections. We will showcase the integration of RDMkit with other knowledge resources in ELIXIR. The talk will also describe RDMkit's contribution and editorial processes as well as its underlying technology and implementation.

Data availability statement

RDMkit content and site infrastructure is open-source and can be found at the following GitHub repositories

- ELIXIR Toolkit Theme: <https://github.com/ELIXIR-Belgium/elixir-toolkit-theme>
- RDMkit content: <https://github.com/elixir-europe/rdmkit>

Underlying and related material

Not applicable.

Author contributions

NF and PA wrote the original draft abstract.

CG and FC supervised the work and reviewed and edited the abstract.

Competing interests

The authors declare that they have no competing interests.

Funding

RDMkit has been developed as part of the Horizon 2020 Project ELIXIR CONVERGE Grant agreement ID: 871075.

The Luxembourg National Data Service (LNDS) is a brand of the Plateforme Nationale d'Échange de Données (PNED G.I.E), an economic interest group established by the Luxembourgish Government.

Acknowledgement

RDMkit's best practice guidelines have been developed in a crowdsourced fashion by an open community contributors listed on <https://rdmkit.elixir-europe.org/contributors>.

References

- [1] Marjan Grootveld, Ellen Leenarts, Sarah Jones, Emilie Hermans, & Eliane Fankhauser. (2018). OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans (1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.1120245>
- [2] Harrow J, Drysdale R, Smith A, Repo S, Lanfear J, Blomberg N. ELIXIR: providing a sustainable infrastructure for life science data at European scale. *Bioinformatics*. 2021 Aug 25;37(16):2506-2511. doi: 10.1093/bioinformatics/btab481. PMID: 34175941; PMCID: PMC8388016.
- [3] RDMkit Community. The Research Data Management toolkit for Life Sciences. <https://rdmkit.elixir-europe.org/>
- [4] Philippe Rocca-Serra, Wei Gu, Vassilios Ioannidis, Tooba Abbassi Daloui, Salvador Capella-Gutierrez, Ishwar Chandramouliswaran, Andrea Splendiani, Tony Burdett, Robert T. Giessmann, David Henderson, Dominique Batista, Allyson Lister, Ibrahim Emam, Yojana Gadiya, Lucas Giovanni, Egon Willighagen, Chris Evelo, Alasdair J. G. Gray, Philip Gribbon, ... the FAIR Cookbook Recipes' Authors. (2022). The FAIR Cookbook - the essential resource for and by FAIR doers (1.0). Zenodo. <https://doi.org/10.5281/zenodo.7156792>
- [5] Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M; FAIRsharing Community. FAIRsharing as a community approach to standards, repositories and policies. *Nat Biotechnol*. 2019 Apr;37(4):358-367. doi: 10.1038/s41587-019-0080-8. PMID: 30940948; PMCID: PMC6785156.
- [6] Niall Beard, Finn Bacall, Aleksandra Nenadic, Milo Thurston, Carole A Goble, Susanna-Assunta Sansone, Teresa K Attwood, TeSS: a platform for discovering life-science training opportunities, *Bioinformatics*, Volume 36, Issue 10, May 2020, Pages 3290–3291, <https://doi.org/10.1093/bioinformatics/btaa047>
- [7] Ison J, Ienasescu H, Chmura P, Rydza E, Ménager H, Kalaš M, Schwämmle V, Grüning B, Beard N, Lopez R, Duvaud S, Stockinger H, Persson B, Vařeková RS, Raček T, Vondrášek J, Peterson H, Salumets A, Jonassen I, Hooft R, Nyrönen T, Valencia A, Capella S, Gelpí J, Zambelli F, Savakis B, Leskošek B, Rapacki K, Blanchet C, Jimenez R, Oliveira A, Vriend G, Collin O, van Helden J, Løngreen P, Brunak S. The bio.tools registry of software tools and data resources for the life sciences. *Genome Biol*. 2019 Aug 12;20(1):164. doi: 10.1186/s13059-019-1772-6. PMID: 31405382; PMCID: PMC6691543.
- [8] Pergl, R., Hooft, R., Suchánek, M., Knaisl, V. and Slifka, J., 2019. "Data Stewardship Wizard": A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning. *Data Science Journal*, 18(1), p.59. DOI: <http://doi.org/10.5334/dsj-2019-059>

Open Science Best Practices in Data Science and Artificial Intelligence

Ekaterina Borisova¹[\[https://orcid.org/0000-0002-3447-9860\]](https://orcid.org/0000-0002-3447-9860),
Raia Abu Ahmad¹[\[https://orcid.org/0009-0004-8720-0116\]](https://orcid.org/0009-0004-8720-0116), and
Georg Rehm¹[\[https://orcid.org/0000-0002-7800-1893\]](https://orcid.org/0000-0002-7800-1893)

¹Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany

Abstract: In the past years, scientific research in Data Science and Artificial Intelligence has witnessed vast progress. The number of published papers and digital objects (e.g., data, code, models) is growing exponentially. However, not all research artefacts fulfill the criteria of being findable, accessible, interoperable and reusable (FAIR), contributing to a rather low level of reproducibility of experimental findings reported in scholarly publications and to the reproducibility crisis. In this paper, we focus on Data Science and Artificial Intelligence Open Science best practices, i.e., a set of recommendations that eventually contribute to the management and development of digital artefacts that are as FAIR as possible. While several guidelines exist, we add best practices for the FAIR collection, processing, storing and sharing of scholarly findings via Research Knowledge Graphs. The final list of recommendations will be available on the NFDI4DS website as an interactive web application.

Keywords: FAIR, Reproducible Research, Open Science

1 Introduction

The past years have seen rapid progress in the fields of Data Science (DS), Machine Learning (ML) and Artificial Intelligence (AI) with neural methods becoming the state-of-the-art in a wide range of research areas such as natural language processing and computer vision. Contemporary computational methods usually consist of code, ML models and data used for their training and evaluation. With an ever-increasing amount of newly appearing ML techniques and datasets, the question of how to make digital objects (especially code, data, models, software) *findable*, *accessible*, *interoperable* and *reusable* (FAIR) [1] to ensure the transparency and reproducibility of research results has never been more relevant and important than today [2], [3].

Recent studies demonstrated that scientific ML and AI pipelines often lack fine-grained documentation. Details on model hyperparameters, data pre-processing steps, evaluation metrics, dependencies, train/test splits, biases, annotation procedures, etc. are either only documented partially or not at all [4], [5]. The absence of this information complicates the validation, replication and improvement of previous findings, i.e., it significantly hinders scientific progress. Furthermore, it is quite common that code,

software and (meta)data are missing or not specified or cited at all in scientific papers [6]. Yousuf et al. [6] show that only about 30% of Computer Science papers from arXiv include links to source code. Since academic literature and search engines are the main data discovery sources for researchers [7], links between digital objects and publications are crucial to guarantee accessibility. In addition, code, models or data are not always publicly available due to privacy, legal, ethical, commercial or copyright restrictions (e.g., medical data, Generative Pre-trained Transformer, GPT, [8] models developed by OpenAI [9], etc.). This factor also contributes to the challenge of FAIR research. All of the aforementioned issues contribute to the *reproducibility crisis* [2], [3] because it is getting increasingly difficult – in many cases it is already impossible – to reuse results and to reproduce state-of-the-art methods.

Reproducibility concerns gave rise to a series of workshops [10]–[15], checklists [16]–[19] and a handbook [20] on FAIR scientific research. Moreover, the availability of digital objects has become a common criterion for evaluating paper submissions at conferences (e.g., NAACL [21], ACL [22]) and in journals (e.g., Nature [23]). However, despite the proposed measures, digital artefacts still tend to be published with incomplete descriptions of their provenance, quality or dependencies [6], [24]. As an extreme example, OpenAI’s technical report on GPT-4 [25] does not provide details on the model’s architecture, dataset construction method or training procedure. This lack of information affects not only research replication but also leads to ethical concerns since it is impossible to verify the use/misuse of personal data during the model training (e.g., private messages). It is essential to encourage scholars and companies to use and develop open-source ML models (such as BLOOM [26]).

2 Open Science Best Practices in Data Science and Artificial Intelligence

To promote the idea of reproducible research, we propose *Open Science best practices especially geared towards Data Science and Artificial Intelligence research*, i.e., recommendations for ensuring a FAIR lifecycle of digital objects. The best practices were collected and summarised based on previous developments in research data and software management (e.g., Gebru et al. [27], Lamprecht et al. [28], Barker et al. [29], Pineau et al. [17], Rogers et al. [18], Dodge et al. [19], Rehm [30], NeurIPS 2021 Paper Checklist Guidelines [16], etc.). It is worth noting that our recommendations are developed primarily for scholars who plan to publish their research, we encourage credibility and findability in science. The core of our best practices constitutes topics ranging from the FAIR collection and processing of data to the distribution, validation and maintenance of (meta)data, code, models and software. For example, our recommendations for code, models and software distribution include but are not limited to:

- Make your code, models or software publicly available. Publish them in an appropriate, recognised and trusted [31] repository. We encourage the use of open source and open access repositories which guarantee the persistent identification (e.g., DOI, PID), long-term availability and authenticity protection of digital artefacts (e.g., Software Heritage [32]);
- Publish code, models or software with rich metadata using an appropriate metadata format;

- Make sure the code can be run out of the box (time and machine independent). Make use of Docker [33] containers and eventually consider publishing them through a community platform such as European Language Grid (ELG) [34].

In addition, the proposed best practices have guidelines for transparent management and sharing of scientific artefacts via Research Knowledge Graphs (RKGs) [35] such as the Open Research Knowledge Graph (ORKG) [36] or the Semantic Scholar Academic Graph (S2AG) [37]. RKGs allow the representation of scientific results and contributions through structured, semantically rich, interlinked knowledge graphs. RKGs are aimed to establish an efficient search across scholarly findings so that researchers can gain an overview of recent developments and are able to compare their findings. In this sense, it has become increasingly crucial that scientific resources are stored, shared, harvested and processed in a FAIR way. For instance, we recommend researchers to label their contributions (e.g., research problem, objective, method, etc.) in the \LaTeX file using the SciKGT eX [38] package to allow the automatic extraction and import of this metadata into RKGs. To the best of our knowledge, the topic of RKGs is not covered by existing resources such as The Turing Way handbook [20] or checklists for reproducible ML and AI research [16]–[19].

To improve user experience and user adoption, especially with regard to junior researchers, our recommendations are aligned with the typical research timeline associated with the development of scientific articles and split into the following four phases:

1. *Before starting the research*
2. *During the research*
3. *Paper submission*
4. *Paper publication*

The DS and AI Open Science best practices we collected are presented in the form of an interactive Streamlit application [39] (see Figure 1). The recommendations will be made available on the NFDI4DS website [40].

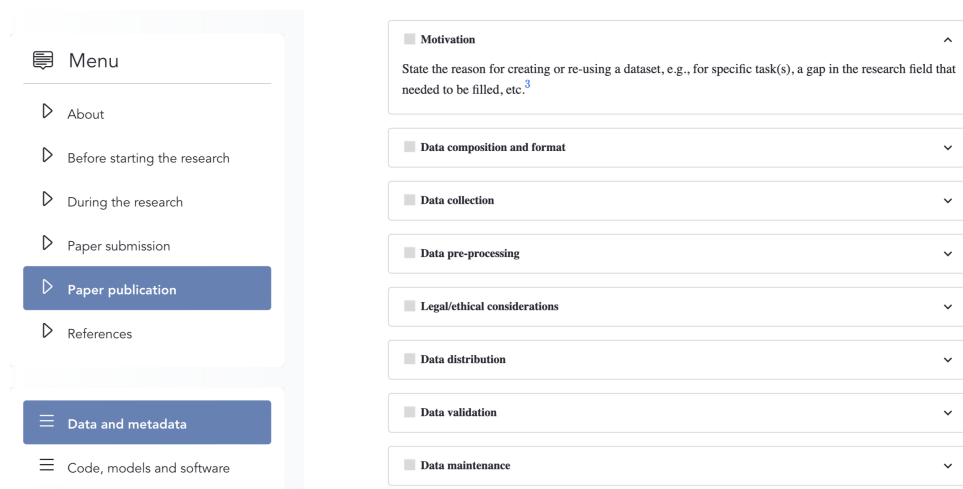


Figure 1. The Open Science Best Practices in Data Science and AI Streamlit application

Data availability statement

No data is shared in the paper.

Author contributions

Georg Rehm conceived the original idea. Ekaterina Borisova and Raia Abu Ahmad performed the research activity planning and execution. Ekaterina Borisova, Raia Abu Ahmad and Georg Rehm wrote this paper.

Competing interests

The authors declare that they have no competing interests.

Funding

This publication was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS) as part of the non-profit association National Research Data Infrastructure (NFDI e.V.). The NFDI is funded by the Federal Republic of Germany and its states. The paper received funding through the German Research Foundation (DFG) project NFDI4DS (no. 460234259). The authors wish to thank both for funding and support. A special thanks goes to all institutions and actors engaging for the association and its goals.

References

- [1] M. Wilkinson, M. Dumontier, I. Aalbersberg, *et al.*, "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data*, vol. 3, no. 160018, Mar. 2016. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
- [2] A. Belz, S. Agarwal, A. Shimorina, and E. Reiter, "A Systematic Review of Reproducibility Research in Natural Language Processing," in *Proc. of the 16th Conf. of the Europ. Chap. of the Assoc. for Comput. Ling.*, Association for Computational Linguistics, May 2021, pp. 381–393. DOI: [10.18653/v1/2021.eacl-main.29](https://doi.org/10.18653/v1/2021.eacl-main.29).
- [3] M. Hutson, "Artificial Intelligence Faces Reproducibility Crisis," *Science*, vol. 359, no. 6377, pp. 725–726, Feb. 2018. DOI: [10.1126/science.359.6377.725](https://doi.org/10.1126/science.359.6377.725).
- [4] F. D. Maurizio, C. Paolo, and J. Dietmar, "Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches," in *Proc. of the 13th Assoc. for Comput. Machinery (ACM) Conf. on Recomm. Systems*, Copenhagen, Denmark: Association for Computing Machinery, Sep. 2019, pp. 101–109. DOI: [10.1145/3298689.3347058](https://doi.org/10.1145/3298689.3347058).
- [5] L. Rupprecht, J. C. Davis, C. Arnold, Y. Gur, and D. Bhagwat, "Improving Reproducibility of Data Science Pipelines Through Transparent Provenance Capture," in *Proc. VLDB Endow.*, VLDB Endowment, Aug. 2020, pp. 3354–3368. DOI: [10.14778/3415478.3415556](https://doi.org/10.14778/3415478.3415556).
- [6] R. B. Yousuf, S. Biswas, K. K. Kaushal, *et al.*, "Lessons from Deep Learning Applied to Scholarly Information Extraction: What Works, What Doesn't, and Future Directions," 2022. arXiv: [2207.04029](https://arxiv.org/abs/2207.04029) [cs.IR].
- [7] K. Gregory, P. Groth, A. Scharnhorst, and S. Wyatt, "Lost or Found? Discovering Data Needed for Research," *Harvard Data Science Review*, vol. 2, no. 2, May 2020. DOI: [10.1162/99608f92.e38165eb](https://doi.org/10.1162/99608f92.e38165eb).
- [8] A. Radford and K. Narasimhan, "Improving Language Understanding by Generative Pre-Training," 2018.
- [9] OpenAI. "OpenAI." Accessed: 2023-04-09. (2023), [Online]. Available: <https://openai.com>.

- [10] A. Lucic, M. Bleeker, S. Bhargav, *et al.*, "Towards Reproducible Machine Learning Research in Natural Language Processing," in *Proc. of the 60th Annual Meeting of the Assoc. for Comput. Ling.: Tutorial Abstracts*, Association for Computational Linguistics, Jun. 2022, pp. 7–11. DOI: [10.18653/v1/2022.acl-tutorials.2](https://doi.org/10.18653/v1/2022.acl-tutorials.2).
- [11] GO FAIR. "M4M Workshop." Accessed: 2023-04-08. (2018), [Online]. Available: <https://www.go-fair.org/events/m4m-workshop/>.
- [12] GO FAIR. "M4M #2: Preclinical trials + M4M #3: Funders." Accessed: 2023-04-08. (2019), [Online]. Available: <https://www.go-fair.org/events/m4m-2-preclinical-trials-m4m-3-funders/>.
- [13] GO FAIR. "The Second GO FAIR Workshop for the German Research Community." Accessed: 2023-04-08. (2018), [Online]. Available: <https://www.go-fair.org/2018/10/08/on-the-road-to-fair/>.
- [14] GO FAIR. "The 3rd Germany GOes FAIR Workshop for the German Research Community." Accessed: 2023-04-08. (2019), [Online]. Available: <https://www.go-fair.org/2019/06/04/3rd-germany-goes-fair-workshop-report/>.
- [15] OpenAIRE and RDA Europe and FAIRsFAIR and EOSC-hub. "Services to Support FAIR Data." Accessed: 2023-04-08. (2019), [Online]. Available: <https://eosc-portal.eu/events/workshop-series-services-support-fair-data>.
- [16] NeurIPS. "NeurIPS 2021 Paper Checklist Guidelines." Accessed: 2023-04-08. (2021), [Online]. Available: <https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist>.
- [17] J. Pineau, P. Vincent-Lamarre, K. Sinha, *et al.*, "Improving Reproducibility in Machine Learning Research (A Report from the Neurips 2019 Reproducibility Program)," *Journal of Machine Learning Research*, vol. 22, no. 164, pp. 1–20, Jan. 2021.
- [18] A. Rogers, T. Baldwin, and K. Leins, "Just What Do You Think You're Doing, Dave? A Checklist for Responsible Data Use in NLP," in *Findings of the Assoc. for Comput. Ling.: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4821–4833. DOI: [10.18653/v1/2021.findings-emnlp.414](https://doi.org/10.18653/v1/2021.findings-emnlp.414).
- [19] J. Dodge, S. Gururangan, D. Card, R. Schwartz, and N. A. Smith, "Show Your Work: Improved Reporting of Experimental Results," in *Proc. of the 2019 Conf. on Empirical Methods in NLP and the 9th Intern. Joint Conf. on NLP (EMNLP-IJCNLP)*, Association for Computational Linguistics, Nov. 2019, pp. 2185–2194. DOI: [10.18653/v1/D19-1224](https://doi.org/10.18653/v1/D19-1224).
- [20] The Turing Way Community. "The Turing Way: A Handbook for Reproducible, Ethical and Collaborative Research (1.0.2)." Accessed: 2023-04-08. (2021), [Online]. Available: <https://doi.org/10.5281/zenodo.7625728>.
- [21] NAACL. "NAACL 2021 Reproducibility Checklist." Accessed: 2023-04-08. (2021), [Online]. Available: <https://2021.naacl.org/calls/reproducibility-checklist/>.
- [22] ACL. "ARR Responsible NLP Research Checklist." Accessed: 2023-04-08. (2021), [Online]. Available: <https://aclrollingreview.org/responsibleNLPresearch/>.
- [23] Nature. "Reporting Standards and Availability of Data, Materials, Code and Protocols." Accessed: 2023-04-08. (2023), [Online]. Available: <https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards>.
- [24] A. Spirling, "Why Open-source Generative AI Models are an Ethical Way Forward for Science," *Nature*, vol. 616, no. 413, Apr. 2023. DOI: [10.1038/d41586-023-01295-4](https://doi.org/10.1038/d41586-023-01295-4).
- [25] OpenAI, "GPT-4 Technical Report," 2023. arXiv: [2303.08774](https://arxiv.org/abs/2303.08774) [cs.CL].
- [26] T. L. Scao, A. Fan, C. Akiki, *et al.*, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," 2023. arXiv: [2211.05100](https://arxiv.org/abs/2211.05100) [cs.CL].

- [27] T. Gebru, J. Morgenstern, B. Vecchione, *et al.*, "Datasheets for Datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021. DOI: [10.1145/3458723](https://doi.org/10.1145/3458723).
- [28] A. L. Lamprecht, L. Garcia, M. Kuzak, *et al.*, "Towards FAIR Principles for Research Software," *Data Science*, vol. 3, no. 1, pp. 37–59, Jun. 2020. DOI: [10.3233/DS-190026](https://doi.org/10.3233/DS-190026).
- [29] M. Barker, N. P. Chue Hong, D. S. Katz, *et al.*, "Introducing the FAIR Principles for Research Software," *Scientific Data*, vol. 9, no. 622, Oct. 2022. DOI: [10.1038/s41597-022-01710-x](https://doi.org/10.1038/s41597-022-01710-x).
- [30] G. Rehm, "The Language Resource Life Cycle: Towards a Generic Model for Creating, Maintaining, Using and Distributing Language Resources," in *Proc. of the 10th Intern. Conf. on Lang. Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), May 2016, pp. 2450–2454.
- [31] D. Lin, J. Crabtree, I. Dillo, *et al.*, "The TRUST Principles for Digital Repositories," *Scientific Data*, vol. 7, Dec. 2020. DOI: [10.1038/s41597-020-0486-7](https://doi.org/10.1038/s41597-020-0486-7).
- [32] Software Heritage. "Software Heritage." Accessed: 2023-07-06. (2023), [Online]. Available: <https://zenodo.org>.
- [33] Docker. "Docker." Accessed: 2023-04-14. (2023), [Online]. Available: <https://www.docker.com>.
- [34] ELG. "European Language Grid." Accessed: 2023-04-14. (2023), [Online]. Available: <https://live.european-language-grid.eu>.
- [35] S. Auer, V. Kovtun, M. Prinz, A. Kasprzik, M. Stocker, and M. E. Vidal, "Towards a Knowledge Graph for Science," in *Proc. of the 8th Intern. Conf. on Web Intelligence, Mining and Semantics*, ser. WIMS '18, Novi Sad, Serbia: Association for Computing Machinery, Jun. 2018, pp. 1–6. DOI: [10.1145/3227609.3227689](https://doi.org/10.1145/3227609.3227689).
- [36] M. Y. Jaradeh, A. Oelen, K. E. Farfar, *et al.*, "Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge," in *Proc. of the 10th Intern. Conf. on Knowledge Capture*, Marina Del Rey, CA, USA: Association for Computing Machinery, Sep. 2019, pp. 243–246. DOI: [10.1145/3360901.3364435](https://doi.org/10.1145/3360901.3364435).
- [37] R. M. Kinney, C. Anastasiades, R. Authur, *et al.*, "The Semantic Scholar Open Data Platform," 2023. arXiv: [2301.10140](https://arxiv.org/abs/2301.10140) [cs.DL].
- [38] C. Bless, I. Baimuratov, and O. Karras, "SciKGT_{EX} – A L^AT_EX Package to Semantically Annotate Contributions in Scientific Publications," 2023. arXiv: [2304.05327](https://arxiv.org/abs/2304.05327) [cs.DL].
- [39] Streamlit. "Streamlit." Accessed: 2023-04-13. (2023), [Online]. Available: <https://streamlit.io>.
- [40] NFDI4DS. "NFDI4DataScience." Accessed: 2023-04-12. (2023), [Online]. Available: <https://www.nfdi4datascience.de>.

On the Design and Implementation of Easy Access to External Spatiotemporal Datasets in NFDI

Christian Beilschmidt¹[\[https://orcid.org/0009-0001-6297-0921\]](https://orcid.org/0009-0001-6297-0921)

Dominik Brandenstein²[\[https://orcid.org/0009-0002-1901-9935\]](https://orcid.org/0009-0002-1901-9935)

Johannes Dröner¹[\[https://orcid.org/0009-0003-9629-2844\]](https://orcid.org/0009-0003-9629-2844)

Nikolaus Glombiewski²[\[https://orcid.org/0000-0003-2876-3918\]](https://orcid.org/0000-0003-2876-3918)

Michael Mattig¹[\[https://orcid.org/0009-0006-1893-5391\]](https://orcid.org/0009-0006-1893-5391)

Bernhard Seeger^{1,2}[\[https://orcid.org/0000-0002-9362-153X\]](https://orcid.org/0000-0002-9362-153X)

¹Geo Engine GmbH, Am Kornacker 68, 35041 Marburg, Germany

²University of Marburg, Dept. of Mathematics and Computer Science,
Hans-Meerwein-Str. 6, 35032 Marburg, Germany

Abstract: n.a.

Keywords: Spatiotemporal data access, Workflow platform, Geo Engine

Across many scientific domains, the ability to process large amounts of heterogeneous spatiotemporal data from various sources is crucial for solving challenging research questions. For example, researchers in NFDI4Biodiversity [1] must combine observational data with satellite images to correlate biodiversity loss with climate change variables.

In general, large datasets are not available on the system (called consumer) where the processing is performed, but first have to be retrieved from one or multiple external systems (called providers) that offer a corresponding service. Moreover, a consumer is often unaware of the datasets the providers offer. Ideally, a provider follows FAIR principles [2] and thus supports mechanisms to simplify data exchange. However, in practice, multiple providers with valuable datasets are not as FAIR as desired or lack spatiotemporal-specific support for data exchange. Instead of improving each potential provider at the source, we propose an intermediary spatiotemporal data exchange layer (SDExL) that helps simplify data exchange so that domain experts can easily access valuable data with little technical know-how.

Based on practical experience and guided by the FAIR principles, in the following, we postulate four requirements for building an SDExL (Figure 1). Then, we discuss two reference implementations within Geo Engine [3], a flexible analytical processing platform for spatiotemporal data used in projects like NFDI4Biodiversity and FAIR Data Spaces [4].

The following three-step process of an SDExL summarizes the general communication between consumers and providers.

SDExL: Required and optional steps

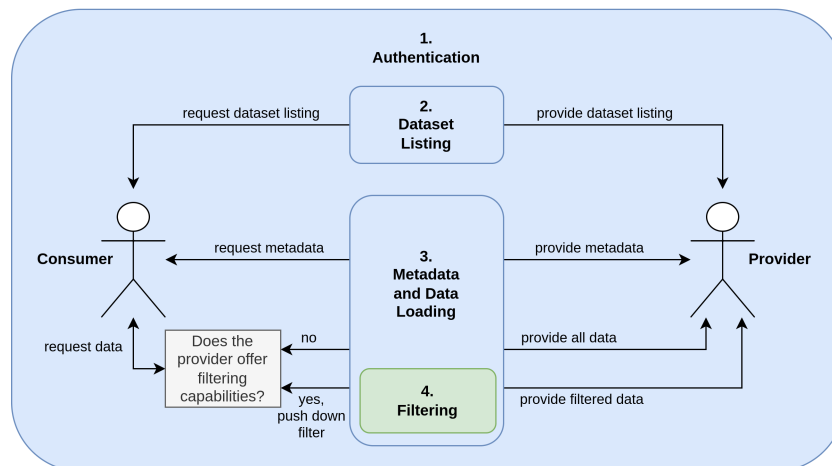


Figure 1. Overview of the spatiotemporal data exchange layer (SDExL).

1. First, users of a consumer have to authenticate themselves in the provider to gain access to datasets. Afterwards, SDExL abstracts away all access control mechanisms until the system's authorization policy requires user input. Furthermore, the provider needs to be uniquely identified in the consumer through an identifier.
2. Next, the users can request a list of all datasets they can access. SDExL translates a list of unique dataset IDs of the provider into a list of unique dataset IDs for the consumer.
3. The users select one of the datasets from their list in a two-step approach. First, SDExL requests any available metadata for the desired dataset. Second, based on the metadata, the actual dataset is requested and delivered to the consumer.

Each step in the process is a strict requirement for implementing an SDExL. The fourth optional requirement minimizes processing time and thus increases the overall usability.

4. At any point, SDExL leverages the filtering capabilities of the provider such that the consumer receives only the data required for the scientific task at hand.

Geo Engine implements multiple connectors for interacting with providers, e.g., standard services based on WCS [5] and STAC [6], following the SDExL process. We now discuss two specialized connectors in detail.

Aruna Object Storage (AOS) [7] is a FAIR cloud-based storage platform developed in NFDI4Biodiversity. As an object storage, AOS has limited spatiotemporal processing capabilities but offers scientific data objects to users. The four requirements are realized as follows.

1. Users of AOS create a key for authorization and pass it to Geo Engine. This key is attached to each request of Geo Engine.
2. Dataset listings are natively available through AOS.
3. AOS provides for each dataset a metaobject containing the loading information.
4. AOS offers labels for datasets, e.g., species names. Geo Engine leverages these labels as a filter to only return a subset of datasets in listings and subsequently to reduce loading of the actual data.

GBIF [8] offers open access to the largest observation database about life on Earth, with over 2 billion records. Many scientific applications use GBIF, enrich the data with domain-specific data, and offer the resulting data products in the open-source database system PostgreSQL [9] with its powerful spatial extension PostGIS [10]. For such a setting, our four requirements are realized as follows.

1. User credentials are required to connect to a PostgreSQL database and access its data records.
2. Dataset listings are retrieved through SQL queries processed by PostgreSQL. E.g., a list of species can be retrieved with a "SELECT UNIQUE species. . ." query, serving as identifiers for occurrences of a species. In order to impose restrictions on data visibility, it is possible to utilize PostgreSQL's role-based access control mechanisms.
3. The PostGIS extension of PostgreSQL offers the required metadata (e.g., spatial reference system) for loading data into Geo Engine. Then, Geo Engine uses this metadata as input for the PostgreSQL GDAL driver [11] to load the actual data.
4. The SQL query interface of PostgreSQL supports accessing datasets and many means of filtering data records. For example, when Geo Engine only requires data within a specific spatial bounding box, the GDAL driver leverages PostgreSQL's filtering capabilities by adding a filter condition to the SQL query.

Finally, Geo Engine can manage datasets internally, either uploaded by an administrator or by the user. These datasets can be combined with datasets provided by SDExL through spatiotemporal processing operators in Geo Engine. The processing steps are stored as a graph of operators, which serves as a provenance workflow with a unique ID. Since workflows in Geo Engine can be accessed like a dataset through standard OGC [12] protocols, Geo Engine is a provider that fulfills requirements 1-3. In addition, it offers spatiotemporal filtering capabilities (requirement 4) that consumers can leverage. Thus, it fulfills our four requirements. By acting as a facilitator for data exchange, Geo Engine is a service that helps researchers solve difficult research questions while ensuring interoperability within the overall spatiotemporal processing landscape.

For future work, we plan to connect Geo Engine to many more data spaces which adhere to our postulated requirements. For example, connecting the Data and Information Access Services [13] will make the satellite data from the Copernicus mission available. Moreover, we will implement SDExL for time series database systems in the project FAIR Data Spaces. Finally, Geo Engine will specify SDExL as a self-describing service in a federated GAIA-X [14] catalog making it available to a large user community, including the one of NFDI [15].

Data availability statement

This submission is not based on data.

Underlying and related material

The source code of Geo Engine is publicly available on GitHub: <https://github.com/geo-engine/geoengine>.

Author contributions

Nikolaus Glombiewski wrote the initial draft (CRediT ID: 43ebbd94-98b4-42f1-866b-c930ce-f228ca). All authors reviewed and edited it to create the final draft (CRediT ID: d3aead86-f2a2-47f7-bb99-79de6421164d).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partially funded by the German Research Foundation DFG under the grant agreement number 442032008 (NFDI4Biodiversity). The project is part of NFDI, the National Research Data Infrastructure Programme in Germany. This work was partially funded by the BMBF project FAIR Data Spaces (FAIRDS10). This work was partially funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under grant numbers O3EUPHE069 and 50EE2303B.

References

- [1] "NFDI4Biodiversity." (2023), [Online]. Available: <https://www.nfdi4biodiversity.org/> (visited on 04/25/2023).
- [2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [3] C. Beilschmidt, J. Dröner, M. Mattig, P. Schweitzer, and B. Seeger, "Geo engine: Workflow-backed geo data portals," in *BTW 2023*, Bonn: Gesellschaft für Informatik e.V., 2023, pp. 837–849, ISBN: 978-3-88579-725-8.
- [4] "FAIR Data Spaces." (2023), [Online]. Available: <https://www.nfdi.de/fair-data-spaces/> (visited on 04/25/2023).
- [5] "Web Coverage Service." (2023), [Online]. Available: <https://www.ogc.org/standard/wcs/> (visited on 04/25/2023).
- [6] "SpatioTemporal Asset Catalogs." (2023), [Online]. Available: <https://stacspec.org/> (visited on 04/25/2023).
- [7] "Aruna Object Storage." (2023), [Online]. Available: <https://www.uni-giessen.de/de/fbz/fb08/Inst/bioinformatik/software/aruna> (visited on 04/25/2023).
- [8] "Global Biodiversity Information Facility." (2023), [Online]. Available: <https://www.gbif.org/> (visited on 04/25/2023).
- [9] "PostgreSQL." (2023), [Online]. Available: <https://www.postgresql.org/> (visited on 04/25/2023).
- [10] "PostGIS." (2023), [Online]. Available: <https://postgis.net/> (visited on 04/25/2023).
- [11] "GDAL PostgreSQL Driver." (2023), [Online]. Available: <https://gdal.org/drivers/vector/pg.html> (visited on 04/25/2023).
- [12] "Open Geospatial Consortium." (2023), [Online]. Available: <https://www.ogc.org/> (visited on 04/25/2023).
- [13] "Data and Information Access Services." (2023), [Online]. Available: <https://www.copernicus.eu/en/access-data/dias> (visited on 04/25/2023).

- [14] "The Gaia-X Ecosystem - A Sovereign Data Infrastructure for Europe." (2023), [Online]. Available: <https://www.bmwk.de/Redaktion/EN/Dossier/gaia-x.html> (visited on 04/25/2023).
- [15] "NFDI - Nationale Forschungsdateninfrastruktur." (2023), [Online]. Available: <https://www.nfdi.de/> (visited on 04/25/2023).

The Aruna Object Storage

A distributed multi cloud object storage system for scientific data management

Marius Alfred Dieckmann¹‡[\[https://orcid.org/0000-0001-5130-546X\]](https://orcid.org/0000-0001-5130-546X), Sebastian Beyvers¹‡[\[https://orcid.org/0000-0002-9747-7096\]](https://orcid.org/0000-0002-9747-7096), Jannis Hochmuth¹‡[\[https://orcid.org/0009-0004-4382-4760\]](https://orcid.org/0009-0004-4382-4760), Anna Rehm²[\[https://orcid.org/0009-0000-5063-486X\]](https://orcid.org/0009-0000-5063-486X), Frank Förster^{1,3}[\[https://orcid.org/0000-0003-4166-5423\]](https://orcid.org/0000-0003-4166-5423), and Alexander Goesmann¹[\[https://orcid.org/0000-0002-7086-2568\]](https://orcid.org/0000-0002-7086-2568)

¹Bioinformatics and Systems Biology, Justus Liebig University, Giessen, Germany

²Algorithmic Bioinformatics, Justus Liebig University, Giessen, Germany

³Bioinformatics Core Facility, Justus Liebig University, Giessen, Germany

‡ Authors with equal contributions

Abstract: The exponential growth of scientific data has led to an increasing demand for effective data management and storage solutions. Academic computing infrastructures are often fragmented, which can make it challenging for researchers to leverage cloud-native principles and modern data analysis tools. To address this challenge, a new distributed storage platform called Aruna Object Storage (AOS) was developed. AOS is a cloud-native, scalable, and domain-agnostic object storage system that provides an S3-compatible interface for a variety of data analysis tools like Apache Spark, TensorFlow, and Pandas. The system uses an underlying distributed NewSQL database to manage detailed information about its resources and can be deployed across multiple data centers for geo-redundancy. AOS is designed to support modern DataOps practices, including the adoption of FAIR principles. Resources in AOS are organized into Objects, Datasets, Collections and Projects, which represent relations of data objects. Additionally, these can be further annotated with key-value pairs called Labels and Hooks to provide additional information about the data. The system's event-driven architecture makes it easy to automate actions and enforce data validation checks, significantly improving accessibility and reproducibility of scientific results. AOS is open source and freely available via <https://aruna-storage.org>.

Keywords: data management, storage, FAIR, multi-cloud, cloud-native, data mesh

1 Introduction

In recent years, significant progress has been made in the field of information technology, resulting in decreasing costs for data processing and data storage [1]. At the same time, the volume and importance of data has increased dramatically. Cloud computing infrastructures have gained popularity and are now widely used for data analysis in the commercial sector. This has led to the development of a variety of novel tools

and frameworks. However, many research communities have been slow to adopt cloud computing and have missed out on many of the benefits of these infrastructures and frameworks, partly due to the fragmented landscape of academic computing infrastructures. As part of the NFDI4Biodiversity and NFDI4Microbiota consortia, it was therefore decided to build a new storage platform that integrates the heterogeneous offerings into a common system and allows researchers to manage their data similarly to FAIR Digital Objects [2]. Our goal is to help scientists use cloud-native principles to improve the quality of their data while simplifying data access and analysis procedures.

2 Goals

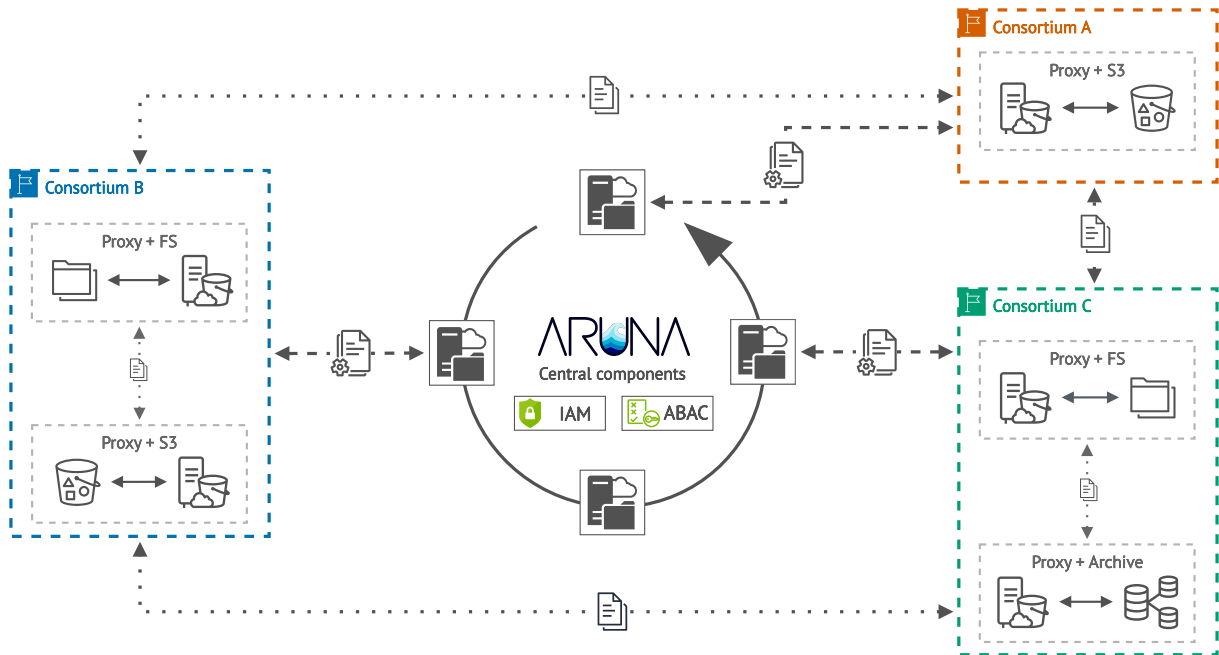
Our project aims to create a cloud-native, geo-redundant, scalable, and domain-agnostic object storage based data mesh system for scientists [3]. It aims to leverage the heterogeneous computing and storage infrastructures of the scientific community enhanced by additional data management capabilities. In addition, it will support researchers using modern data analysis tools such as Apache Spark [4], Nextflow [5] or Pandas [6] by providing an S3-compatible interface. It is also intended to enable researchers to adopt modern DataOps practices for better and more efficient data handling throughout the data life cycle. Consequently, the system will facilitate the application of the FAIR principles [7]. Primary access to the system should be API-based, supporting OAuth2 and OIDC for authentication and Attribute Based Access Control (ABAC) for authorization [8].

3 Results

Based on these goals, the Aruna Object Storage (AOS) was developed. It is implemented in Rust and provides multiple access methods for end users, such as a gRPC [9] and JSON-over-REST API, as well as pre-built client libraries for multiple programming languages. The system uses an underlying distributed NewSQL database to manage detailed information about its resources (Fig. 1). The database can be deployed across multiple data centers and scaled horizontally to keep pace with the growth of the data stored. Data submitted by users is stored using data proxies, which provide an S3-compatible API with additional functionality to abstract from existing storage infrastructures. This allows a variety of different academic computing and storage providers to be integrated into the system, enabling easy and automated offsite backups and site-local caches, while allowing participants to retain full data sovereignty.

3.1 Objects and higher-level resources

All data uploaded and stored by users is stored as an Object, represented as a sequence of bytes without any semantic information. Once uploaded, these Objects are immutable. Updates create new Objects that reference the original Object, resulting in a history of changes. Objects are organized into Collections and optional Datasets. A Dataset consists of closely related Objects and is used to combine data and metadata for easier access and organization. Collections and Projects, on the other hand, contain a set of Objects and Datasets that represent a scoped view of the data. Collections, Datasets and Projects can be snapshotted, capturing the current state and providing a persistent, versioned identifier. This allows other researchers to accurately reproduce results based on a specific version, while allowing for continuous modification of the current data. All resources and their relationships form a directed acyclic graph (DAG) with Projects as roots and Objects as leaves (Fig. 2).



IAM: Identity and access management, ABAC: Attribute based access control, S3: Simple storage service, FS: File System

Figure 1. Schematic overview of centralised and decentralised AOS components. The centralised AOS components handle authentication and authorisation by integrating existing IAM providers in combination with user-specific attributes (ABAC). The central components also provide a registry with meta-descriptions and locality information making records discoverable. The decentralised components consist of data proxy applications that expose existing data structures via a common S3 interface and enable data exchange and caching in a peer to peer network within and between participants.

3.2 Data life cycle and events

Resources, and in particular Objects, have their own lifecycle, expressed by states. These can be used to apply certain checks to the data before it is made available. Using webhooks, a user could request that a particular piece of data conforms to a particular standard. After the data is uploaded, the system checks it against the specified standard before making it available. Only if the validation is successful is it made available, otherwise it is placed in an error state that only the user in question can see, without being visible to external users by default. In addition, all state changes emit events that can trigger additional external automation.

4 Discussion

AOS is a modern data management and storage system that has a number of distinct advantages over other tools. First and foremost, it aims to provide a cloud-native solution that is scalable and flexible to connect to a variety of cloud and non-cloud compute and storage infrastructures. By providing an S3-compatible interface, AOS is able to interact with a variety of modern data analysis tools such as Pandas or the Hadoop ecosystem. Its additional features such as labels, webhooks and events make it easy for scientists to adopt DataOps practices to make their work more efficient. Using a

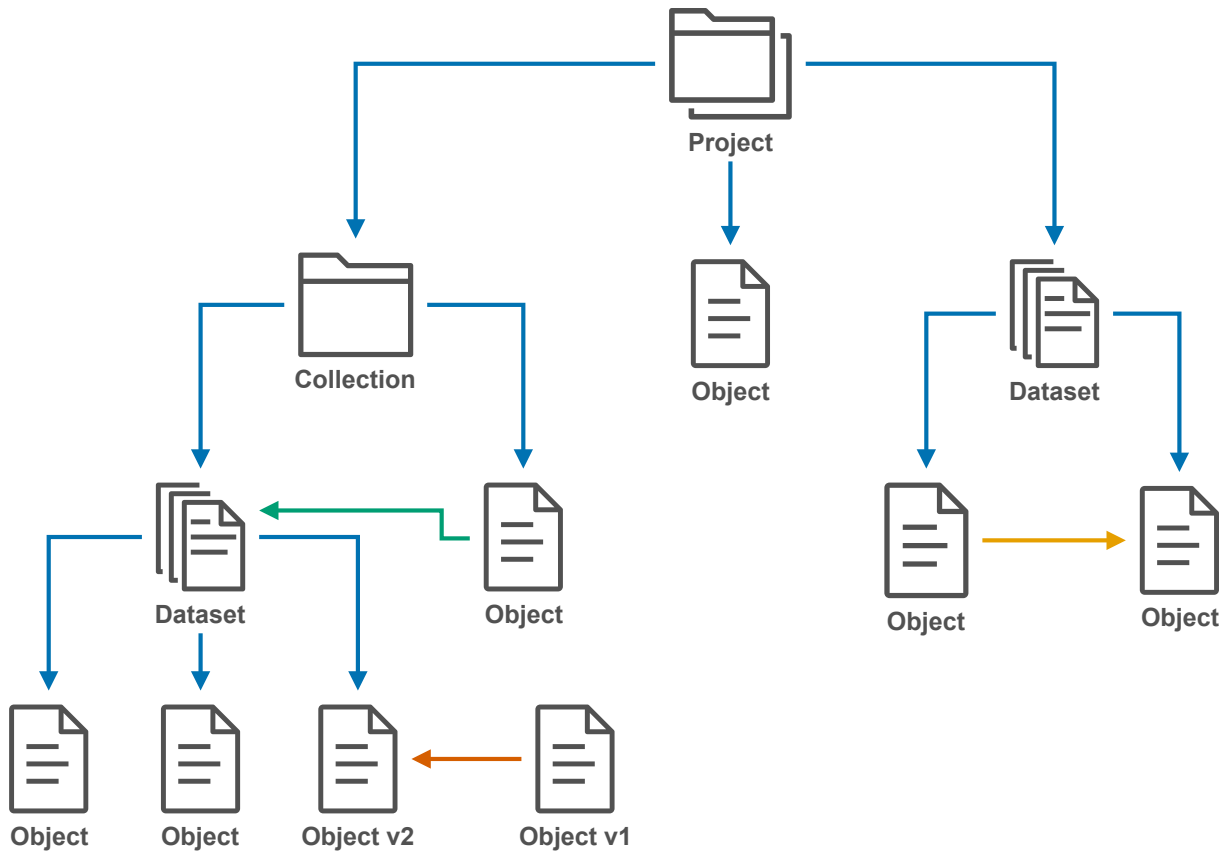


Figure 2. Hierarchical structure of AOS resources. Resources form a directed acyclic graph of *belongs to* relationships (blue) with Projects as roots and Objects as leaves. Resources can also describe horizontal *version* relationships (orange), *data/metadata* relationships (yellow) or even custom user-defined relationships (green).

distributed database that can be deployed across multiple data centers in Germany reduces the risk of the system being unavailable due to problems in a single data center. AOS is open source and freely available; more information, including documentation and source code, is available at <https://aruna-storage.org>

Author contributions

Conceptualization: MAD, SB, JH, FF; **Data curation:** JH; **Formal analysis:** MAD, SB; **Funding acquisition:** AG; **Investigation:** MAD, SB, JH, AR, FF; **Methodology:** MAD, SB, JH, FF; **Project administration:** MAD, FF, AG; **Resources:** MAD, SB, JH, FF, AG; **Software:** MAD, SB, JH, FF; **Supervision:** MAD, FF, AG; **Validation:** MAD, SB, JH, FF; **Visualization:** MAD, SB, JH, AR, FF; **Writing – original draft:** MAD, SB, FF; **Writing – review & editing:** MAD, SB, JH, AR, FF, AG

Competing interests

The authors declare that they have no competing interests.

Funding

MAD is funded by the German Research Foundation DFG under the National Research Data Infrastructure–NFDI4BioDiversity(NFDI 5/1)–442032008. AR and FF are funded by the German Research Foundation DFG under the National Research Data Infrastructure–NFDI4Microbiota(NFDI 28/1)–460129525. SB and JH are funded via FAIR Data Spaces by the German Federal Ministry of Education and Research BMBF (grant FAIRDS08).

Acknowledgements

Thanks to Dr. Karina Brinkrolf for proofreading the manuscript. We thank Amazon Web Services for distributing the icons used in Fig. 1 & 2 under the CC-BY-ND 2.0 license.

References

- [1] C. L. Borgman, "The conundrum of sharing research data," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 6, pp. 1059–1078, 2012. DOI: <https://doi.org/10.1002/asi.22634>.
- [2] K. De Smedt, D. Koureas, and P. Wittenburg, "Fair digital objects for science: From data pieces to actionable knowledge units," *Publications*, vol. 8, no. 2, p. 21, 2020. DOI: <https://doi.org/10.3390/publications8020021>.
- [3] I. A. Machado, C. Costa, and M. Y. Santos, "Data mesh: Concepts and principles of a paradigm shift in data architectures," *Procedia Computer Science*, vol. 196, pp. 263–271, 2022. DOI: <https://doi.org/10.1016/j.procs.2021.12.013>.
- [4] S. Salloum, R. Dautov, X. Chen, P. X. Peng, and J. Z. Huang, "Big data analytics on apache spark," *International Journal of Data Science and Analytics*, vol. 1, pp. 145–164, 2016. DOI: <https://doi.org/10.1007/s41060-016-0027-9>.
- [5] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature biotechnology*, vol. 35, no. 4, pp. 316–319, 2017. DOI: <https://doi.org/10.1038/nbt.3820>.
- [6] W. McKinney *et al.*, "Pandas: A foundational python library for data analysis and statistics," *Python for high performance and scientific computing*, vol. 14, no. 9, pp. 1–9, 2011.
- [7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016. DOI: <https://doi.org/10.1038/sdata.2016.18>.
- [8] E. Yuan and J. Tong, "Attributed based access control (abac) for web services," in *IEEE International Conference on Web Services (ICWS'05)*, IEEE, 2005. DOI: <https://doi.org/10.1109/ICWS.2005.25>.
- [9] K. Indrasiri and D. Kuruppu, *gRPC: up and running: building cloud native applications with Go and Java for Docker and Kubernetes*. O'Reilly Media, 2020, ISBN: 9781492058335.

RSpace + iRODS

A scalable, flexible and versatile solution that facilitates data and metadata interoperability and is suitable for deployment in conjunction with a wide range of e-infrastructures and Research Commons

Rory Macneil¹<https://orcid.org/0000-0002-8429-096X> and Terrell Russell²

¹ Research Space, Edinburgh, Scotland

² Renaissance Computing Institute, University of North Carolina at Chapel Hill, USA

Abstract.

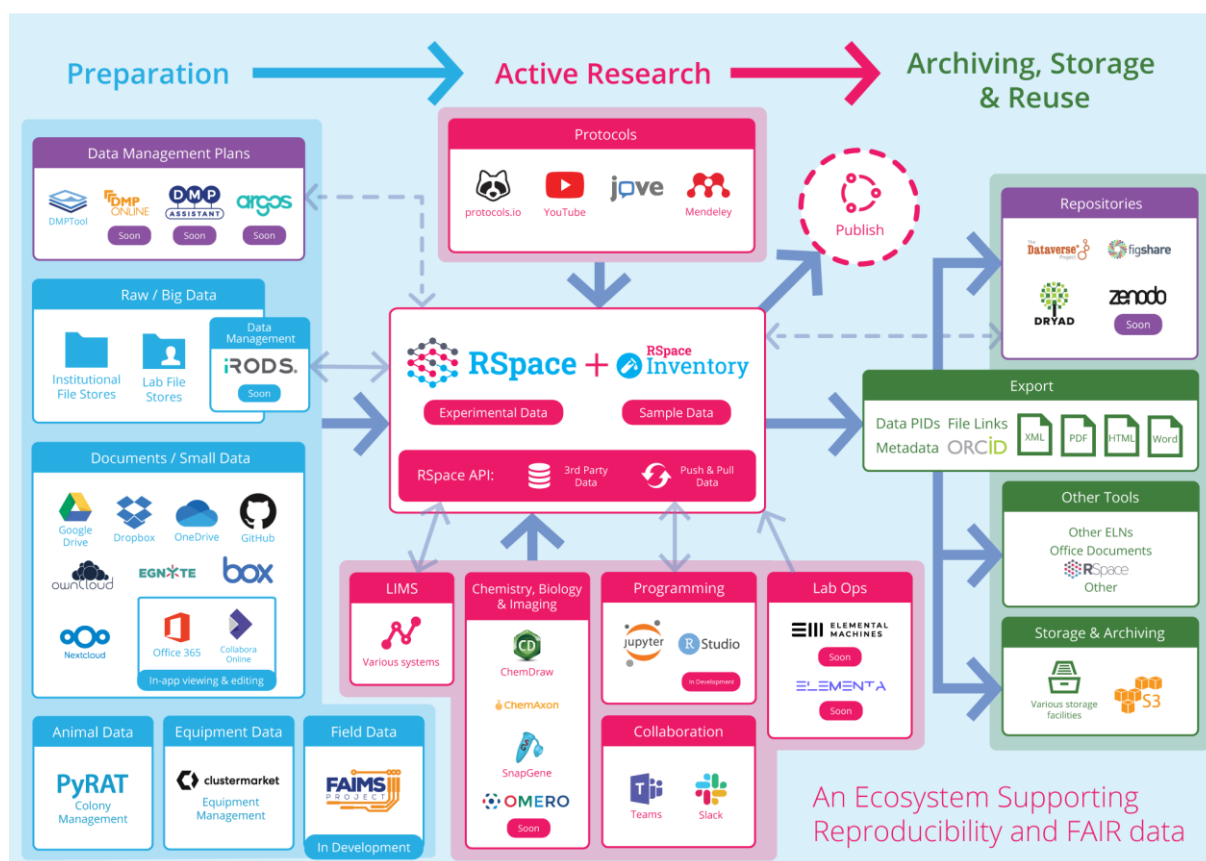
Research infrastructures enabling scalable sharing of data are proliferating, at the institutional, project/consortium and national/international levels. Many of these are domain-specific, but there also is a growing focus on Research Commons that enable general data sharing across domains. Examples include the EUDAT Collaborative Data Infrastructure in Europe, the Gakunin RDM platform in Japan, and the Research Commons planned by Canada's Digital Research Alliance (DRA).

In some cases, such as Gakunin RDM, Research Commons are architected as an integrated set of complementary services, but in most existing and planned Research Commons a disparate set of unconnected services is offered. This paper describes the integration between [RSpace](#), an active content management digital research platform, and [iRODS](#), a policy-driven data management platform, and explains how it is designed to serve as a flexible connecting component that ties together other services that make up a Research Commons.

The paper discusses how, by tying together otherwise unconnected services, inclusion of RSpace + iRODS in Research Commons enables streamlined flows of data and metadata between services, enhancing FAIR principles. This will be illustrated by considering inclusion of RSpace + iRODS in the two specific examples of Canada's proposed Research Commons and the EUDAT Collaborative Data Infrastructure. In both cases the interaction between RSpace, iRODS, and individual services provided by the Commons will be discussed, and a comparison will be made between the two Commons and the benefits derived from the inclusion of RSpace + iRODS.

RSpace and iRODS

As noted, RSpace is a digital content management platform designed to interoperate with other research tools and resources. RSpace and the ecosystem and workflows it supports is captured in this graphic:

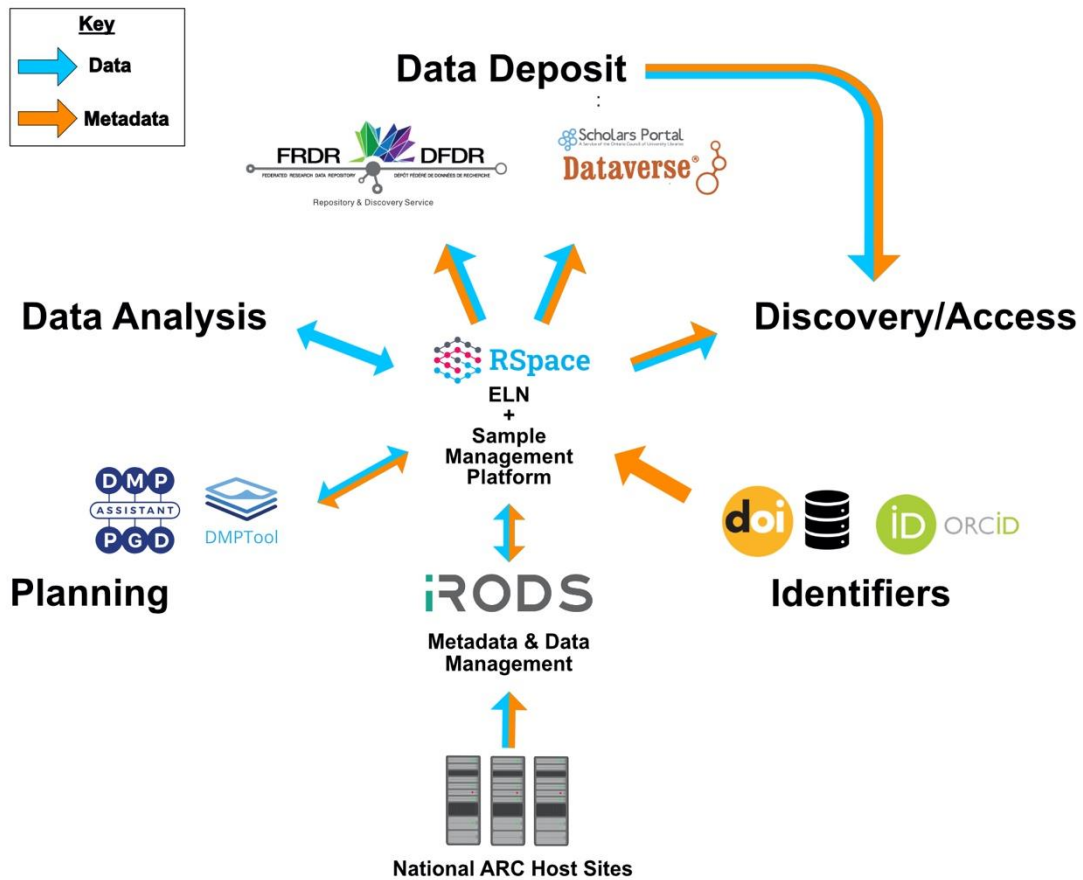


The Integrated Rule-Oriented Data System (iRODS) is open source data management software used by research, commercial, and governmental organizations worldwide. It provides data virtualization with a logical namespace that abstracts the disparate physical storage systems that house the files, a metadata catalog to hold system and user-defined metadata to aid in discovery, and a policy engine to apply admin and user-defined rules to data, based on the metadata in the system. This separation of concerns affords a flexibility to satisfy a wide variety of use cases and scenarios across many diverse domains.

The existing integration between RSpace and the iRODS logical namespace solves the 'broken links' problem by ensuring that, even if files in external resources linked to from RSpace move location the integrity of the link is maintained, thus enhancing the quality and durability of the research record and discoverability. A second phase of the integration is currently being implemented and will be available at the time of the conference in September. This second phase will enable exposure of PIDS and other kinds of metadata from RSpace to iRODS, and association of the metadata with data and metadata in the entire ecosystem of files to which iRODS has access, which could include all the files held by a university, in a large project or collaboration, or a national facility or facilities.

Canada's Research Commons

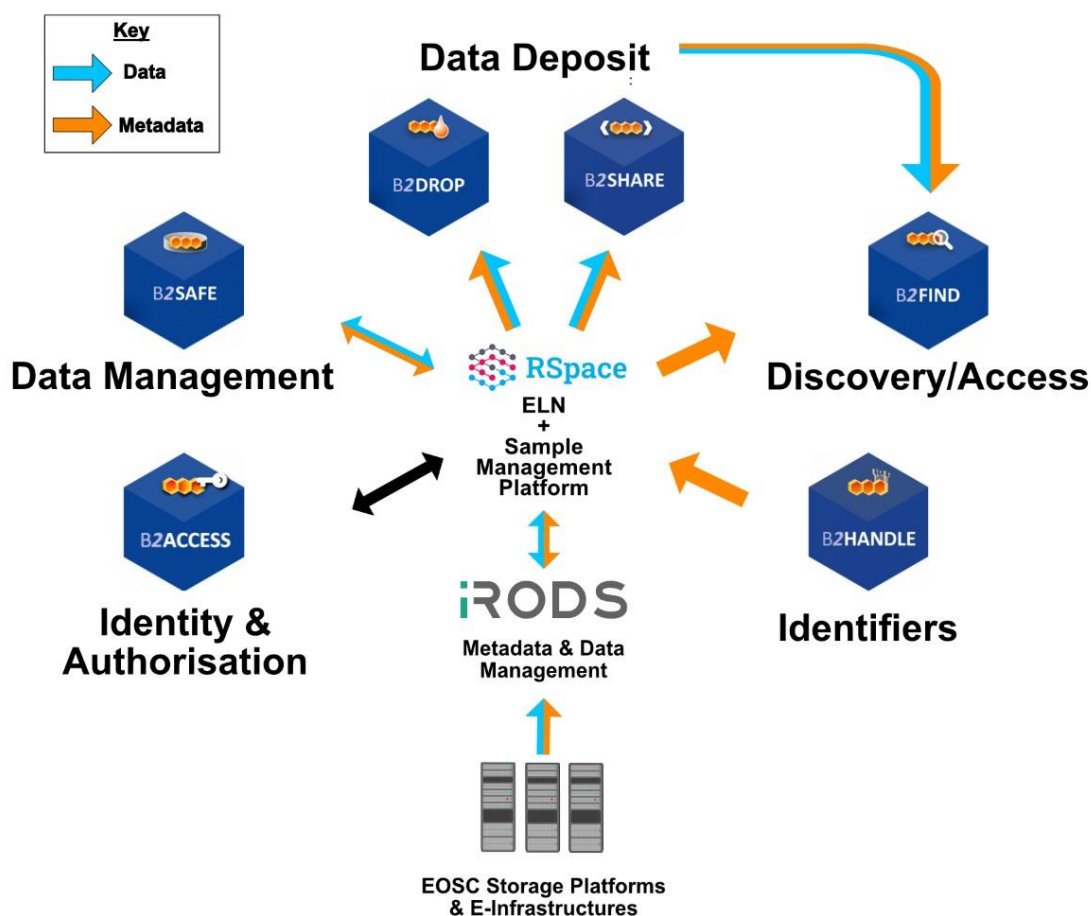
A vision for Canada's Research Commons is set out in this presentation by Mark Leggott, International Director of Canada's Digital Research Alliance, [Digital Research Alliance of Canada as a Research Commons](#). The Commons is to be built around five existing services provided by the DRA: a data storage underpinning in five national data centers; a data management planning tool/service; a national PID delivery service that includes ORCID and DataCite DOIS, and two national data repositories. The Research Commons will be delivered in partnership with Canadian universities and for use by the universities and their researchers. This graphic depicts the vision for inclusion of RSpace + iRODS as a central connecting and catalyzing force in the Commons:



EUDAT Collaborative Data Infrastructure

The [EUDAT CDI](#) consists of six services: B2Share is a data repository; B2Handle is a PID delivery service, B2Drop is a file sharing service; B2Access is an identity and authorization service; B2Safe, a kind of omnibus data policy/management service and actually *is* iRODS; B2Find is a data discovery service. As is the case in the envisioned DRA Commons, the six services are delivered as separate services, although plans are underway to enable some limited integration between the services.

This graphic depicts the vision for inclusion of RSpace + iRODS as a central connecting and catalyzing force in the EUDAT CDI:



Concluding comments

It is interesting and important that all three Commons referred to in this paper, the EUDAT CDI, the DRA Research Commons, and the Gakunin RDM, are built around resources that are (a) pre-existing and (b) intended for general use generally by researchers in multiple domains. The Commons have adopted, and in some cases adapted, existing resources, but they have not attempted to reinvent the wheel by duplicating existing resources, and the Commons are all intended for use in multiple research domains.

A second point of interest is that the EUDAT CDI and the DRA Research Commons both include a similar common set of core services relating to provision of data storage, data management planning, repositories, and provision of PIDS.

We believe that use of pre-existing, generalist tools both represents best practice and is the approach best suited to creating truly scalable RDM solutions. The fact that all three of the Commons mentioned have adopted this approach also facilitates incorporation of RSpace into the Commons, since RSpace already integrates with many of the resources included in both the EUDAT CDI and the DRA Research Commons, including data storage facilities, data management planning tools, data repositories, and PIDS delivery services.

It is, moreover, encouraging that both the EUDAT CDI and the DRA Research Commons share these resources as core components. The inclusion of RSpace + iRODS as an integrating mechanism into these two early attempts to create Research Commons could as described transform these Commons into a more powerful, comprehensive service. It could also provide a model for the development of similar initiatives in other jurisdictions, and because of the shared use of common elements, a series of Commons that are themselves interoperable. Thus this paper is also relevant to the Connecting RDM track.

Keywords: Data, Metadata, Interoperability, FAIR Research Commons

Research Data Publication at Large Scale

Thomas Zastrow¹ [<https://orcid.org/0000-0002-2844-1495>] and Nicolas Fabas¹ [<https://orcid.org/0000-0002-2224-0780>]

¹ Max Planck Computing and Data Facility

Introduction

The MPCDF is the high performance computing center of the Max Planck Society. Beside hosting compute clusters, around 300 petabytes of research data are stored at the MPCDF. Many of these datasets have a size of several terabytes up to petabytes and are stored over a heterogeneous landscape of storage systems. The datasets are covering a wide variety of scientific disciplines, many different data formats and access restrictions. For these reasons, it is not possible to offer one centralized data publishing solution for the datasets. Instead, the MPCDF developed a flexible data publishing concept (see Fig. 1), taking into account the challenges mentioned above. In this paper, we will describe the two parts of our data publishing concept: first, the creation and handling of metadata and second, the operating model for data repositories.

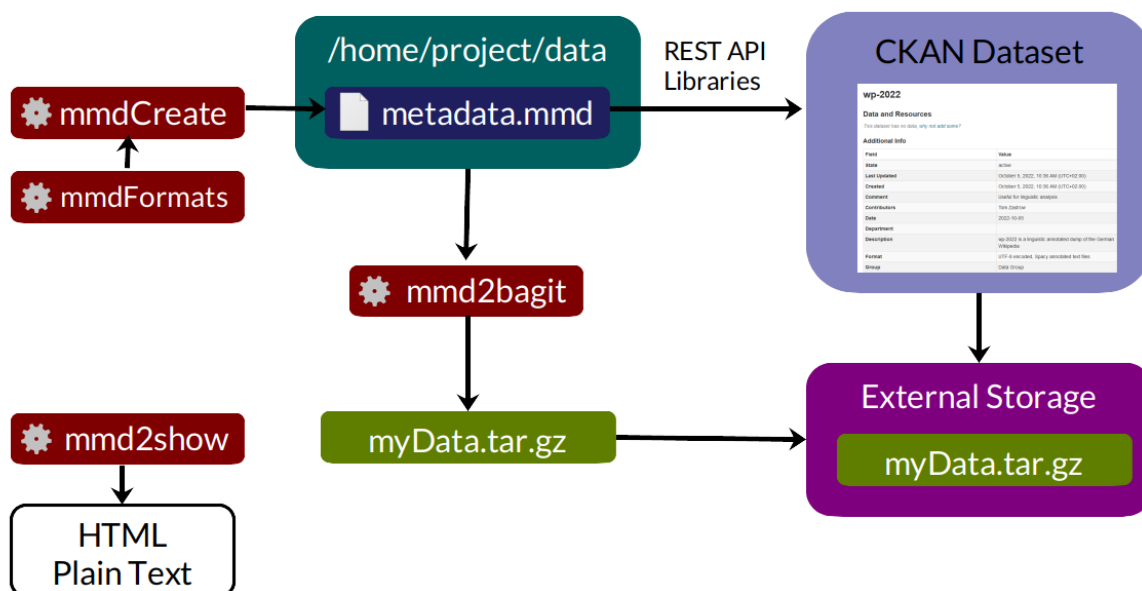


Figure 1: Research Data Publishing Workflow

Metadata Handling: The Metadata Tool Suite

If research data should be published in a data repository, descriptive metadata is necessary. This includes both administrative metadata (Who owns the data? Where is it stored?) as well as content related metadata (What is the data about? Who is allowed to access the data?).

Some of this metadata can be gathered automatically, but some information needs to be added manually. To support the latter type of metadata, the MPCDF developed the "MPCDF Metadata Tool Suite" (MMD Tools)¹. The tool suite contains command line scripts which are flexible to be adapted and extended by the researchers themselves. The created metadata is stored in standardized JSON files, which allows a transformation into other formats as well as importing it into a data repository. Several common metadata formats are supported out of the box: currently, these are Dublin Core², the DataCite Metadata Schema (version 4.4)³ and our own custom format which covers the basic needs of metadata handling at the MPCDF. But the MMD Tools are not restricted to these formats: supported by the MPCDF, researchers can create their own metadata schemata.

In detail, the tool suite contains the following scripts:

- **mmdCreate**: can be used to interactively create a new metadata file or edit an existing one
- **mmdShow**: displays or converts the content of a metadata file
- **mmd2bagit**: creates a self describing BagIt⁴ container out of a given metadata file and the corresponding object data

In addition to the executable scripts, the tool suite contains a Python module with common functionality to be included in individual workflows. Below, a simple example of the stored metadata in Dublin Core format is shown:

```
{
  "id": "c8f6ea3d-c9ce-4cc4-99e1-7d36a7ed014f",
  "mmdFormat": "Dublin Core",
  "Format": "UTF-8 Encoded Text Files",
  "Type": "Annotated Text Corpus",
  "Language": "DE",
  "Title": "The wp-2022 Corpus",
  "Subject": "annotated text corpus, POS tagging, lemmatization",
  "Coverage": "Dump of the German Wikipedia from January 2022",
  "Description": "The corpus contains token, pos and lemma annotations",
  "Creator": "Dr. Thomas Zastrow",
  "Publisher": "---",
  "Contributor": "---",
  "Rights": "Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA) and the GNU Free Documentation License (GFDL)",
  "Source": "https://wikipedia.de",
  "Relation": "---",
  "Date": "2022-01-01"
}
```

¹ The MMD Tools are freely available under Apache 2.0 license and can be downloaded from the MPCDF GitLab. Installation instructions and further documentation can be found here: <https://docs.mpcdf.mpg.de/doc/data/publication/mmd.html>

² <https://www.dublincore.org/specifications/dublin-core/>

³ <https://schema.datacite.org/>

⁴ <https://www.rfc-editor.org/info/rfc8493>

Data Repositories: From Generic to Specific

Data repositories are a common way to publish research data in a FAIR way. At the MPCDF, the heterogeneous kind of the data stored makes it difficult to provide one repository for all the datasets. Therefore, the MPCDF developed a two-step procedure to fulfill the needs of the Max Planck institutes. The basis is a repository software off the shelf, which is installed and initially configured by the MPCDF. After that, the repository is handed over to the institute. Any adaptations and further configurations can then be done by the researchers or IT staff of the responsible institute. The MPCDF is still supporting the ongoing work with practical help in any directions, but, after initialization and configuration, the responsibility lies at the institute.

As there are many (open source) software solutions available which can be used as basis for a data repository, the MPCDF decided to go with the "Comprehensive Knowledge Archive Network" (CKAN)⁵. For our purposes, the CKAN software has several advantages. One key feature is the possibility to store only metadata in the repository and leave the object data where it is: datasets at the MPCDF are often too big to be easily moved around or to get downloaded via a web browser. Another feature is the extensive REST-style API which can be used for automation from nearly any programming language. Built on top of the REST API, wrapper libraries for several programming languages are available. In addition to the basic functionality of the CKAN software, the MPCDF supports the use of some extra plugins and developed specifically a plugin to assign DOIs to a dataset using MPG's DOI service.

Any CKAN instance can store metadata following one or more schemata. These metadata schemata can be harmonized with the metadata created by the MMD Tool suite as mentioned above. By utilizing the repository's API or implementing the available programming libraries, automated metadata and data publishing workflows are possible.

Currently, the MPCDF hosts several individual data repositories for Max Planck institutes. They are covering a range of scientific disciplines from archeology, physics (dark matter research and physics of light) to material research (polymer research). In some use cases, the data repository is part of a bigger (automated) workflow. This includes for example the integration of Jupyter Notebooks as well as the integration of HPC systems for analysing the data.

Conclusion

The combination of the MMD tool suite for metadata handling and CKAN as repository software is an easy and flexible solution for publishing heterogeneous research data. It allows the researchers to quickly and conveniently add metadata to their project's data in the same environment they are using for research. After that, an adapted and highly configurable data repository allows the publication of research data in a FAIR way.

⁵ <https://ckan.org/>

The HMC Information Portal for enhanced metadata collaboration in the Helmholtz FAIR data space

Lucas Kulla¹[\[https://orcid.org/0000-0002-2484-2742\]](https://orcid.org/0000-0002-2484-2742), Jens Bröder²[\[https://orcid.org/0000-0001-7939-226X\]](https://orcid.org/0000-0001-7939-226X), Constanze Curdt³[\[https://orcid.org/0000-0002-9606-9883\]](https://orcid.org/0000-0002-9606-9883), Markus Kubin⁴[\[https://orcid.org/0000-0002-2209-9385\]](https://orcid.org/0000-0002-2209-9385), Helen Kol-
lai⁵[\[https://orcid.org/0000-0003-0214-1336\]](https://orcid.org/0000-0003-0214-1336), Christine Lemster⁶[\[https://orcid.org/0000-0001-7764-1517\]](https://orcid.org/0000-0001-7764-1517), Marco
Nolden¹[\[https://orcid.org/0000-0001-9629-0564\]](https://orcid.org/0000-0001-9629-0564), Kai Schmieder⁷[\[https://orcid.org/0000-0002-1171-1428\]](https://orcid.org/0000-0002-1171-1428), Annika
Strupp⁸[\[https://orcid.org/0000-0002-0070-4337\]](https://orcid.org/0000-0002-0070-4337), Karl-Uwe Stucky⁷[\[https://orcid.org/0000-0002-0065-0762\]](https://orcid.org/0000-0002-0065-0762), Emanuel
Söding⁵[\[https://orcid.org/0000-0002-4467-642X\]](https://orcid.org/0000-0002-4467-642X), Konstantin Pascal Walter⁴[\[https://orcid.org/0009-0003-4693-0932\]](https://orcid.org/0009-0003-4693-0932) and
Arndt Witold⁹[\[https://orcid.org/0000-0002-7713-9647\]](https://orcid.org/0000-0002-7713-9647),

¹ Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany

²Forschungszentrum Jülich GmbH

³GEOMAR Helmholtz Centre for Ocean Research Kiel

⁴Helmholtz-Zentrum Berlin

⁵Helmholtz Centre for Environmental Research UFZ

⁶GEOMAR Helmholtz Centre for Ocean Research Kiel

⁷Karlsruhe Institute of Technology

⁸Forschungszentrum Jülich

⁹German Aerospace Center

The **Helmholtz Metadata Collaboration** (HMC) platform was launched in late 2019 to turn **FAIR** (Findable, Accessible, Interoperable, Reusable) research data into reality within the Helmholtz Association and beyond. The **Information Portal** was initiated to enable the structured **cartography of metadata** and FAIR landscape of Helmholtz, providing information for multi-level decision-making and creating a curated knowledge base for research data managers, scientists and other stakeholders.

Developed through a top-down approach, 18 categories, and associated metadata schemas were defined and aligned by an HMC taskforce. Data curation followed, with resources collected from different domains based on the aligned metadata schema. The Information Portal is a **web application** for capturing **FAIR data practices** across all Helmholtz domains, offering a unified user interface for collecting and exploring results.

Built using **state-of-the-art technologies**, including Python and Docker, the Information Portal leverages GitLab as a database. It offers a public / central read-only version for stakeholders and a personal instance for curation - synchronized to a GitLab repository. Git-based systems offer advantages, such as raw data accessibility, flexible data curation, easy synchronization, and customizable repositories.

The single-page web application is **user-friendly** and developed in multiple iterations for an intuitive and flexible interface. The Information Portal is crucial for creating a **sustainable, distributed, semantically** enriched Helmholtz data space, promoting seamless data sharing and reuse.

Keywords: Helmholtz Metadata Collaboration, Information Portal, FAIR

The Helmholtz Metadata Collaboration (HMC) platform was launched in late 2019. Its mission is to leverage the visibility and reusability of data across the Helmholtz Association and beyond and to turn FAIR (Findable, Accessible, Interoperable, Reusable) [2] research data into reality. HMC operates as a federated collaboration embedded in different Helmholtz research fields, providing a unique opportunity to translate global metadata concepts into practical implementations. HMCs vision is to create a sustainable, distributed, semantically enriched Helmholtz data space that scientists can use to seamlessly share and re-use data in new ways.

To achieve this goal, the development of the Information Portal was initiated. The Information Portal enables structured cartography of metadata and FAIR landscape of Helmholtz and beyond, providing information for multi-level decision-making for different stakeholders, and creating a curated knowledge base for research data managers and scientists.

Developed in a top-down approach, 18 categories and their associated metadata schemas were aligned and defined between the domains by a taskforce of HMC such as repositories, metadata standards and terminologies. Followed by the alignment, data collection and curation were the next steps. Therefore, the landscape of different domains within the Helmholtz collaboration was explored and data for the various resources were collected based on the defined and aligned metadata schema. The Information Portal is a web application to capture FAIR data practices and resources across all Helmholtz. It uses a unified user interface for capturing and exploring the results of the landscape design process in different areas. It is built using state-of-the-art technologies, with a Python and Docker-based system, leveraging the GitLab Backend of the Helmholtz Cloud [1] for the management of contributions, users, and quality control through CI pipelines. Currently, there are two versions available. The first is a public version, intended to be used as a knowledge base for different stakeholders within Helmholtz (<https://informationportal.helmholtz-metadaten.de>). This version features a read-only view, where users can explore the landscape and find useful information for their work. The second is a personal instance, which is intended to be hosted either on an internal server or on a personal device. It is based on the public version but includes all the tools necessary to curate resources. The edited, deleted, or added resources are then synchronized to a pre-defined GitLab repository. Reusing a Git-based system as a database has multiple advantages over a standard, e.g. Postgres database. The first is that the raw data is accessible by all users of GitLab even without access to the Information Portal. Secondly, curation may or may not be done through the Information Portal, as everyone has the option to add data to the GitLab repository in their preferred way. Thirdly, this allows easy synchronization between multiple private Information Portal instances. Lastly, it is not necessary to use the default GitLab repository. To change the destination of truth a single URL needs to be changed.

Creating a private instance is due to the use of docker-compose a straightforward process, regardless if its deployment is on a local machine or on a server. In addition, the single-page web application is built as user-friendly as possible. Developed in multiple iterations with multiple technological approaches, the user interface is as intuitive, flexible, and target oriented as possible.

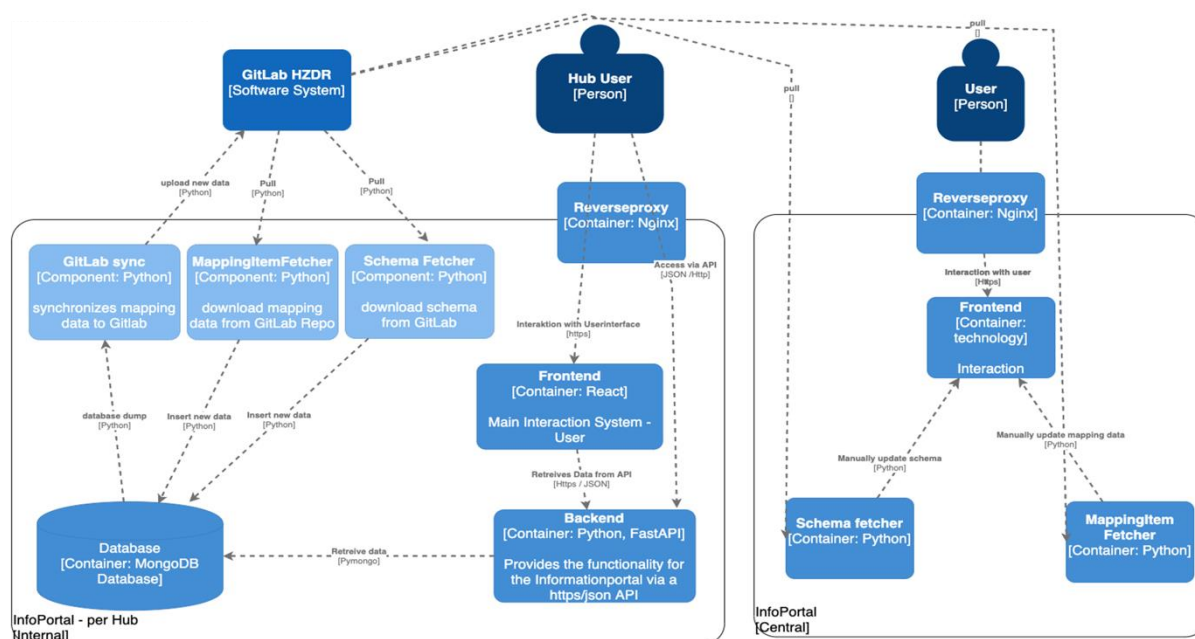


Figure 1. Architecture of the Information Portal

In conclusion, the development of the InformationPortal within HMC has been a crucial step in achieving the goal of creating a sustainable, distributed, semantically enriched Helmholtz data space that scientists can use to seamlessly share and reuse data in new ways. Through a top-down approach and successful collaboration across different units, the Information Portal has been developed utilizing state-of-the-art technologies, providing a centralized resource for metadata and FAIR landscape mapping.

The future outlook of the Information Portal includes several improvements. First, establishing a public contribution and quality assurance process. Secondly, semantic enrichment in the context of Linked Data. The goal is to further utilize the use of data in a way that data is not only seen as independent entries but being seen as a network of entries that are connected to each other.

In conclusion, HMC's pioneering efforts not only empower researchers to unlock the full potential of their and others' data but also pave the way for unprecedented scientific discoveries and collaborations, ultimately driving innovation and shaping the future of research in the Helmholtz Association and beyond.

Author contributions

In the following paragraph, author contributions are summarized following the CRediT taxonomy.

Name	Concep-tual-iza-tion	Data cura-tion	In-vesti-gation	Meth-odol-ogy	Pro-ject ad-min-istra-tion	Soft-ware	Su-per-vi-sion	Writ-ing – origi-nal draft	Writ-ing – re-view & ed-iting
Lucas Kulla	X	X	X	X	X	X		X	
Constanze Curdt	X	X	X	X	X				
Markus Kubin	X	X	X	X	X				X
Helen Kollai	X	X	X	X	X				
Christine Lemster	X	X	X	X	X				
Marco Nolden					X		X		X
Oonagh Mannix					X		X		X
Kai Schmieder	X	X	X	X					
Annika Strupp	X	X	X	X					
Karl-Uwe Stucky	X	X	X	X					X
Emanuel Söding	X	X	X	X	X		X		
Konstantin Pascal Walter	X	X	X	X					
Arndt Witold	X		X		X		X		

Competing interests

The authors declare that they have no competing interests.

Funding

This publication was supported within the hub Health at the German Cancer Research Center by the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative.

References

1. DESY, D. E.-S. (2023). *HIFIS*. Deutsches Elektronen-Synchrotron DESY. Retrieved 26.04.2023 from <https://hifis.net>
2. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., & Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.

Born-fair data projects using cookiecutter templates

Felix Henninger¹[\[https://orcid.org/0000-0002-7730-9511\]](https://orcid.org/0000-0002-7730-9511)

¹ Ludwig-Maximilians-Universität, Germany

Keywords: RDM, Data stewardship, FAIR data principles, Research Software, BERD@NFDI

Implementing research data management best practices and fair principles (Wilkinson et al., 2016) is vital for transparent, reproducible research, as well as efficient, sustainable science that avoids duplication of effort. Scientists can also benefit directly from incorporating data management into their data collection and analysis workflows. However, there is an initial cost to adoption that poses a burden and substantial barrier to entry even to well-intentioned researchers. In our experience in statistical and rdm-focused consulting, this cost increases as a project progresses, with a late-stage conversion being the most costly in terms of resources and energy because a working analysis needs to be adapted in one step. Therefore, we believe that it is useful to adopt best practices for data stewardship early-on in a project, and ideally from the get-go. In this contribution, we present a tool for creating and instantiating project templates that conform to good practices with regard to the data management and analysis projects more generally. In the same vein as “born-open data” (Rouder, 2016) where data is published immediately upon collection, our goal is to establish born-fair datasets that implement proven methods for data stewardship as early on in the research datalifecycle as is feasible. Our aim is to encourage researchers and analysts to incorporate best practices into their workflows from the onset by providing data and analysis templates that implement desirable properties. By adopting these templates, researchers immediately gain access to a number of tools that simplify their work and make it more efficient, while also providing a foundation for increased reproducibility, data documentation through codebooks, as well as metadata for long-term archival. Because a one-size-fits-all approach may not work in practice due to idiosyncrasies of individual research projects, one size may not fit all. For this reason, the templates contain customisation options that researchers can use to tailor the templates to their requirements.

The templates build on the well-established cookiecutter library for the Python programming language (Greenfeld et al., 2022), which we additionally extend to R, a programming language somewhat more common among statisticians and social scientists, thereby creating a cross-platform infrastructure. Both libraries create a project skeleton with a pre-specified directory structure, and include configuration for commonly used tools. Upon template creation, a wizard guides users through a customisation step, allowing them to adapt templates to their needs and catering to the demands of a project at hand.

Owing to the open-source nature of the project and the firmly established and well-documented standard, researchers can easily adapt templates and create their own, to accommodate their specific needs and domain requirements. We hope to foster a community of researchers who share and improve their workflows, and anticipate further uses of the templates for teaching and other purposes.

At CoRDI, we hope to introduce our project to the wider nfdi community and propose it as a lightweight, interoperable and interdisciplinary standard, benefiting all researchers across domains. By streamlining advanced users' workflows, and making reproducible practices more accessible, we aim to enable and facilitate the uptake of rdm across the communities represented there, and build integrations to interoperate with the multitude of services currently under development and in use.

To summarise, we introduce templates for data analysis and archival that researchers can apply themselves, to render possible and encourage better practices during analysis, and prepare data for long-term storage and later re-use. Our hope is to encourage more researchers to adopt rdm best practices more frequently and earlier in projects, demonstrating the value of a more structured workflow and facilitating a shift to FAIR principles more generally

References

1. Greenfeld, A. R., Greenfeld, D. R., Pierzina, R., & Contributors (2022). Cookiecutter (Version 2.1.1) [Computer software]. GitHub. <https://github.com/cookiecutter/cookiecutter/>
2. Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), Article 1. <https://doi.org/10.1038/sdata.2016.18>

Base4NFDI - Basic Services for NFDI

Creating NFDI-wide basic services in a world of specific domains

Sonja Schimmler¹[\[https://orcid.org/0000-0002-8786-7250\]](https://orcid.org/0000-0002-8786-7250), Reinhard
Altenhöner²[\[https://orcid.org/0000-0001-8274-780X\]](https://orcid.org/0000-0001-8274-780X), Lars Bernard³[\[https://orcid.org/0000-0002-3085-7457\]](https://orcid.org/0000-0002-3085-7457),
Juliane Fluck⁴[\[https://orcid.org/0000-0003-1379-7023\]](https://orcid.org/0000-0003-1379-7023), Axel Klinger⁵[\[https://orcid.org/0000-0001-6442-3510\]](https://orcid.org/0000-0001-6442-3510), Sören
Lorenz⁶[\[https://orcid.org/0000-0001-8577-6614\]](https://orcid.org/0000-0001-8577-6614), Brigitte Mathiak⁷[\[https://orcid.org/0000-0003-1793-9615\]](https://orcid.org/0000-0003-1793-9615),
Bernhard Miller⁷[\[https://orcid.org/0000-0002-4385-7245\]](https://orcid.org/0000-0002-4385-7245), Raphael Ritz⁸[\[https://orcid.org/0000-0003-4615-6804\]](https://orcid.org/0000-0003-4615-6804),
Thomas Schörner-Sadenius⁹[\[https://orcid.org/0000-0002-7213-0352\]](https://orcid.org/0000-0002-7213-0352), Alexander
Sczyrba¹⁰[\[https://orcid.org/0000-0002-4405-3847\]](https://orcid.org/0000-0002-4405-3847), and Regine Stein¹¹[\[https://orcid.org/0000-0003-3406-5104\]](https://orcid.org/0000-0003-3406-5104)

¹Fraunhofer Institute for Open Communication Systems (FOKUS), Berlin

²Stiftung Preussischer Kulturbesitz - Staatsbibliothek zu Berlin

³Technische Universität Dresden

⁴ZB Med Information Centre for Life Sciences, Cologne

⁵Technische Informationsbibliothek (TIB), Hannover

⁶GEOMAR Helmholtz-Zentrum für Ozeanforschung, Kiel

⁷GESIS - Leibniz Institute for the Social Sciences, Mannheim and Cologne

⁸Max Planck Computing and Data Facility (MPCDF), Munich

⁹Deutsches Elektronen-Synchrotron, Hamburg

¹⁰Universität Bielefeld

¹¹Georg-August-Universität Göttingen

Abstract: NFDI is a German initiative to set up research data infrastructures across all disciplines. Within NFDI, Base4NFDI is a unique joint effort of all NFDI consortia to develop and deploy NFDI-wide basic services. In this talk, we will provide an overview of Base4NFDI, especially its structures and emerging work program, and inform about ways to participate and contribute ideas for potential basic services.

Keywords: cross-cutting topics, basic services, NFDI, Germany

1 NFDI and Base4NFDI

NFDI¹ is a German initiative to set up and consolidate research data infrastructures across all disciplines, covering Engineering Sciences, Humanities and Social Sciences, Life Sciences, and Natural Sciences. To ensure sustainability, it will integrate national with international activities.

¹www.nfdi.de

Building on activities by domain-specific NFDI consortia, Base4NFDI² represents a cross-disciplinary collaboration. It is a unique joint effort of all 26 NFDI consortia to develop and deploy NFDI-wide basic services. These services will be integrated into the emerging infrastructures at the European level, especially the EOSC. The target group for basic services is the wider NFDI-community and, in particular, operators of community-specific services. The resulting NFDI-wide basic service portfolio will be beneficial for all disciplines by facilitating core tasks in research data management. A service in this context is defined as technical-organisational solution which typically includes storage and computing services, software, processes and workflows, as well as the necessary personnel support for different service desks. A basic service is meant to add value for the consortia and their users typically by bundling existing services. It is characterized by scalability and a model for sustainable operation, thereby requiring a certain degree of technicality. It is meant to deliver effectiveness - measurable by KPIs - over time and regarding usage.

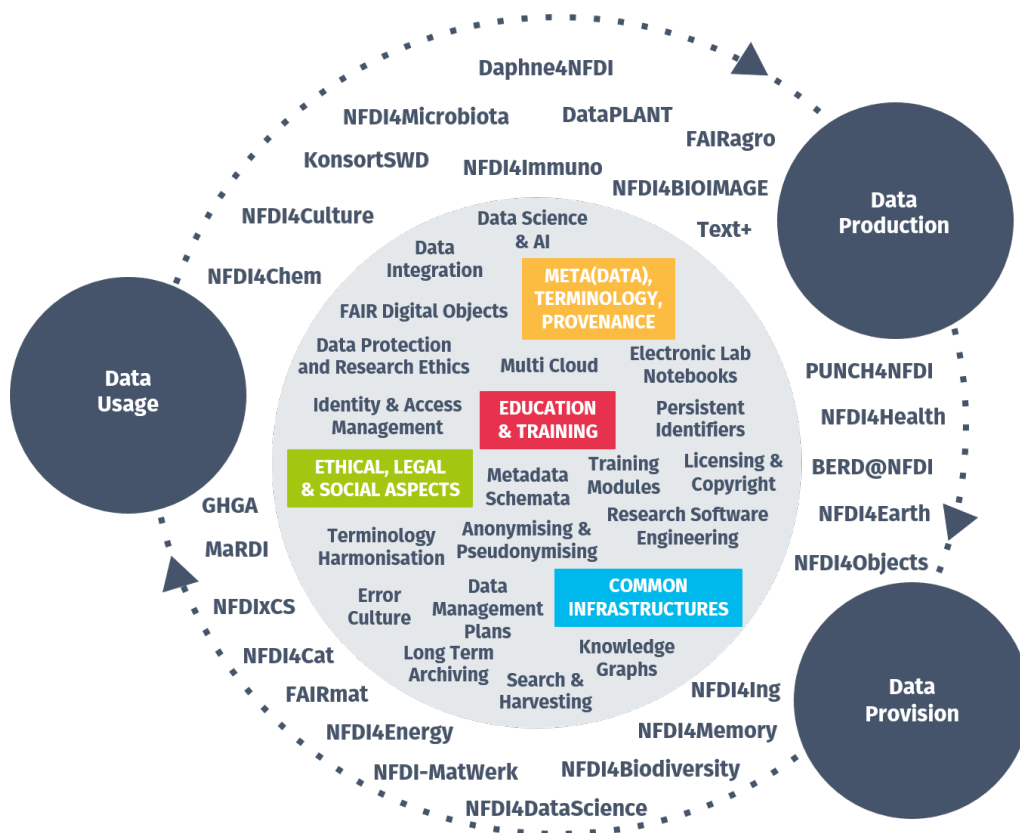


Figure 1. Cross-cutting topics within NFDI

Decisions on basic services are made by all consortia in the bodies of the NFDI Association. To generate proposals for basic services, Base4NFDI will draw on the expertise in NFDI's Sections. They are the loci for exchange between consortia on cross-cutting topics, provide infrastructural and technological expertise in combination with domain knowledge and act as incubators for identifying potential basic services. Potential basic services are thus identified in the Sections, each of which focuses on a topical area: 'Common Infrastructures', 'Education and Training', 'Ethical, Legal and Social Aspects', 'Metadata, Terminologies and Provenance', and the newly established Section 'Industry Engagement'. This broad approach to identifying ideas and needs for basic services reflects, that services relating to research data management must

²www.base4nfdi.de

support a large number of use-cases also including non-technical but organizational ones and thus cover elements like staff for a helpdesk.

After two rounds of proposals in February and May 2023, development will commence with services for Identity and Access Management (IAM), Persistent Identifiers (PID) as well as Terminologies. The IAM Service aims to establish approved identities and organizationally defined access rights across service providers, which will be crucial for seamless data management workflows. The PID Service aims to build on established infrastructures and address challenges such as different levels of maturity in PID implementation across domains. The Terminology Service aims to standardise and harmonise terminology management within NFDI, thereby facilitating consensus-building and interoperability of services across disciplines.

For development, Base4NFDI relies on a three-stage process of 1) initialisation of basic services 2) integration of basic services and 3) ramping-up for operation and becoming part of the NFDI basic service portfolio.

The work program of Base4NFDI is clustered in four Task Areas. Task Areas 'Service requirements, design and development' and 'Service integration and ramping-up for operation' will accompany and support the basic services within the three phases of the process. Task Area 'Service coherence processes and monitoring' will overlook the whole process, and Task Area 'Project governance' will manage the project.

Within this talk, we will give an overview of Base4NFDI, especially its structures and the status of its emerging work program. We will inform about ways to participate and contribute ideas for potential basic services.

Competing Interests

The authors declare that they have no competing interests.

Funding

This joint project received funding by the Deutsche Forschungsgemeinschaft (DFG) – project numbers: 521453681, 521460392, 521462155, 521463400, 521466146, 521471126, 521473512, 521474032, 521475185, 521476232 .

Towards a Research Data Commons in the German National Research Data Infrastructure NFDI : Vision, Governance, Architecture

Michael Diepenbroek¹[\[https://orcid.org/0000-0003-3096-6829\]](https://orcid.org/0000-0003-3096-6829), Ivaylo Kostadinov¹[\[https://orcid.org/0000-0003-4476-6764\]](https://orcid.org/0000-0003-4476-6764), Bernhard Seeger²[\[https://orcid.org/0000-0002-9362-153X\]](https://orcid.org/0000-0002-9362-153X), Frank Oliver Glöckner^{3,4}[\[https://orcid.org/0000-0001-8528-9023\]](https://orcid.org/0000-0001-8528-9023), Marius Dieckmann⁵[\[https://orcid.org/0000-0001-5130-546X\]](https://orcid.org/0000-0001-5130-546X), Alexander Goesmann⁵[\[https://orcid.org/0000-0002-7086-2568\]](https://orcid.org/0000-0002-7086-2568), Barbara Ebert¹[\[https://orcid.org/0000-0003-3328-6693\]](https://orcid.org/0000-0003-3328-6693), Sonja Schimmler^{6,7}[\[https://orcid.org/0000-0002-8786-7250\]](https://orcid.org/0000-0002-8786-7250), York Sure-Vetter⁸[\[https://orcid.org/0000-0002-4522-1099\]](https://orcid.org/0000-0002-4522-1099)

¹ GFBio - German Federation for Biological Data, Germany

² University of Marburg, Germany

³ MARUM - Center for Marine Environmental Sciences, University of Bremen, Germany

⁴ AWI - Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Germany

⁵ Justus Liebig University Giessen, Germany

⁶ Fraunhofer FOKUS, Germany

⁷ Technical University of Berlin, Germany

⁸ NFDI - National Research Data Infrastructure & KIT - Karlsruhe Institute of Technology, Germany

Abstract.

The concept of a "Research Data Commons" (RDC) established itself as an infrastructure ecosystem for science based on open standards and federated resources to facilitate the sharing of research data and services. The consortia of the German National Research Data Infrastructure (NFDI) have identified the collaborative provisioning of resources and services to be of key importance for a functioning and efficient RDC and are leveraging different corresponding measures to establish a sustainable concept in line with international developments.

Keywords: Research Data Commons, Cloud Infrastructure, NFDI, Germany

The concept of a "Research Data Commons" (RDC) established itself as an infrastructure ecosystem for science based on open standards and federated resources to facilitate the sharing of research data and services. The "commons" principle is based on the belief that resources are best used when they are managed in a collaborative and participatory manner [1].

The German National Research Data Infrastructure (NFDI) [2], with 27 consortia [3], identified cross-cutting topics [4] and initiated a number of sections [5] covering and fostering these topics including an RDC.

Internationally, the concept of a "Research Data Commons" has become increasingly important for the development of science infrastructures. Examples include the Australian

Research Data Commons (ARDC) [6], the US National Cancer Institute Research Data Commons (NCI RDC) [7], and, as an important part of the European research and innovation strategy, the European Open Science Cloud (EOSC) [8]. The Global Open Research Commons Interest Group (GORC IG) [9] of the Research Data Alliance (RDA), finally, assembles existing initiatives and heads at more convergence and networking between the various developments.

All of these initiatives aim to provide a unified, open and trusted digital infrastructure for the storage, sharing and reuse of research data. This should enable researchers to integrate data from different sources and thus gain efficiency and new scientific knowledge. In EOSC in particular, structures have already been formed to deal with comparable issues of harmonization of data infrastructures, including implementation projects and working groups that negotiate agreements on relevant fundamental topics (e.g. EOSC Interoperability Framework [10]). The NFDI relates to this central European infrastructure project on several levels: as an overall organization, and via both the consortia and the sections. It is of utmost importance for the development of the NFDI RDC to follow the EOSC principles [11] and to incorporate relevant results from the EOSC technical standardization groups.

The RDC concept is also in line with the goals of the International Data Spaces Association (IDSA) [12], which promotes a virtual data space leveraging existing standards and technologies, as well as governance models. In addition, IDSA supports data sovereignty as a crucial design aspect and proposes a corresponding Reference Architecture Model. The related project FAIR Data Spaces (FAIR DS) [13] aims at the practical concretization of a reference architecture and provides various demonstrators in close cooperation with NFDI. Several NFDI consortia contribute to FAIR DS.

Further, with the beginning of 2023 cross-cutting topics are supported by a new consortium, Base4NFDI [14]. The consortium aims at organizing and evaluating the development of basic services, available at scale, by well-structured workflows embedded into the NFDI governance. Base4NFDI is supposed to support infrastructural supply for potentially all consortia, whereby competing developments and incompatible solutions should be avoided.

The NFDI consortia bring with them a variety of heterogeneous and distributed information infrastructures that, today, are networked only to a limited extent. To establish cross-domain basic services, harmonization within the community of users and providers is indispensable. In doing so, the NFDI builds on decades of experience of its member institutions with the provision of central services. Often, these services are well-tested in single domains, but not harmonized across domains. A central requirement for the NFDI sections is to identify components for common use and to develop proposals on how these can be technically structured and implemented, and which prerequisites are necessary for networking. The sections thereby follow the guiding principle of the NFDI as a whole: existing data and services should be reused as much as possible and intelligently connected.

A specific goal is the NFDI RDC, fostered by the section Common Infrastructures [15]. The current RDC concept is essentially based on a multi-cloud infrastructure that can be used by the consortia and their partners to pool and share data, software, compute resources, and services. Hybrid solutions and edge components are not excluded. The RDC concept envisages various shared cloud services like an application layer with access to high-performance computing, collaborative workspaces, a federated framework for (meta)data integration, persistent identifier, terminology services, data management planning and long-term archiving services as well as a software marketplace (Fig. 1).

Security of data and services is seen as a further important topic. The RDC will be equipped with a common authentication and authorization infrastructure (AAI) and certification services (zero trust). Particularly noteworthy are current works focusing on the development of RDC components performed by the consortia NFDI4Biodiversity [16], NFDI4Microbiota [17],

NFDI4DataScience [18], and NFDI4Earth [19]. The first two pursue an RDC approach as a central concept for their infrastructure development strategy, in particular a distributed core storage infrastructure [20]. In addition, NFDI4Biodiversity is working on effective ways to integrate data and metadata from diverse data providers.

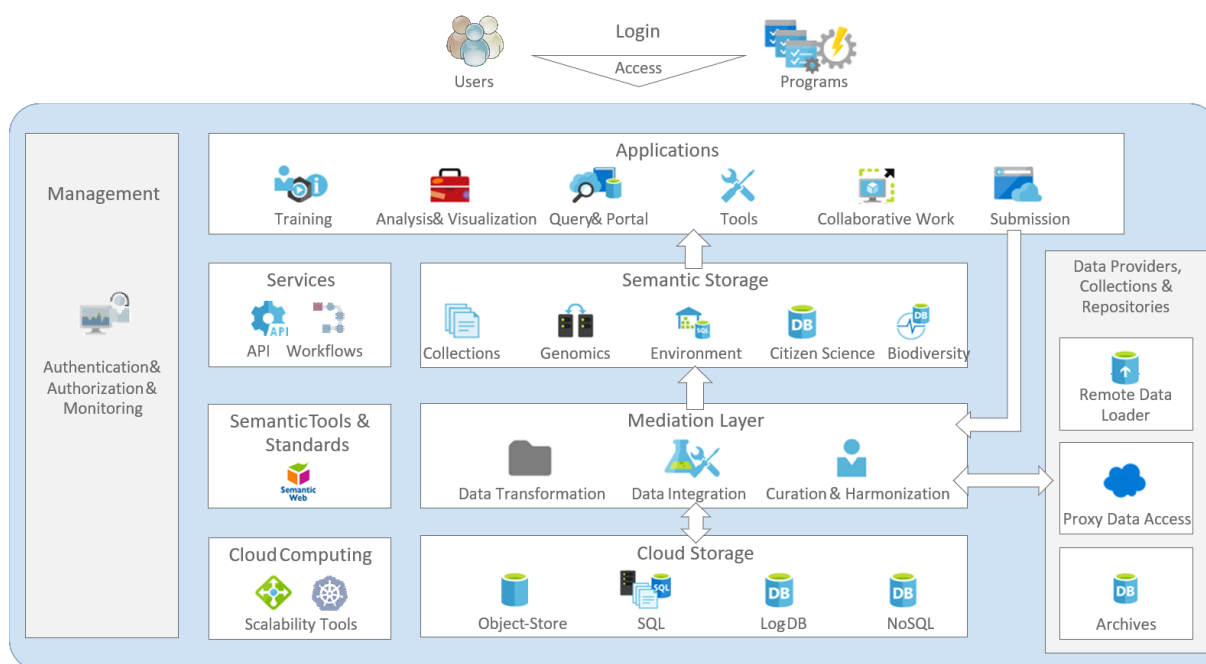


Figure 1. Overview of the RDC architecture, envisioned by the NFDI4Biodiversity and NFDI4Microbiota consortia. Realistically, a final architecture for the entire NFDI will amount to a networking of RDCs resembling a data mesh [21].

Outlook

The NFDI RDC is a holistic infrastructure concept for managing scientific data and services. The RDC implies a federated governance as a foundation of a participative bridge between data and service providers and consumers. It is expected to have a significant effect on the openness and quality of data and suggests a substantial increase in efficiency in the use of existing data. Overall, this provides strong incentives for potential users of NFDI offerings.

The first major challenge is embedding a national RDC into the national and international scientific services landscape, especially with regard to RDC developments in other countries. Within NFDI, in particular in the light of the basic services developments, intensive discussions and negotiations are needed to identify communalities and eventually agree on a common RDC vision and architecture. In view of the prerequisites in the various consortia and general IT developments, an agile development process is expected.

A further challenge is to organize the necessary capabilities and capacities, in particular for building up federated RDC components. Important aspects to address in this respect are compliance and sustainability of supplied services.

Data availability statement

The submission is not based on data.

Underlying and related material

None.

Competing interests

The authors declare that they have no competing interests.

Funding

This publication was written in the context of the work of the association German National Research Data Infrastructure (NFDI), an initiative of the Joint Science Minister Conference of the Federal Republic of Germany and the 16 federal states. The work of the NFDI consortia is funded by the German Research Foundation DFG: NFDI4Biodiversity (grant number 442032008), NFDI4Microbiota (grant number 460129525), NFDI4DataScience (grant number 460234259), NFDI4Earth (grant number 460036893).

Acknowledgement

We thank the dedicated staff and partners who help to shape and support the structures and results presented here.

References

1. Grossman, R.L. Ten lessons for data sharing with a data commons. *Sci Data* 10, 120 (2023). <https://doi.org/10.1038/s41597-023-02029-x>
2. Nationale Forschungsdaten Infrastruktur (NFDI). <https://www.nfdi.de/?lang=en> (2023-04-24)
3. Nationale Forschungsdaten Infrastruktur (NFDI) - Consortia. <https://www.nfdi.de/consortia/?lang=en> (2023-04-24)
4. Glöckner, Frank Oliver, Diepenbroek, Michael, Felden, Janine, Overmann, Jörg, Bonn, Aletta, Gemeinholzer, Birgit, Güntsch, Anton, König-Ries, Birgitta, Seeger, Bernhard, Pollex-Krüger, Annette, Fluck, Juliane, Pigeot, Iris, Kirsten Toralf, Mühlhaus, Timo, Wolf, Christof, Heinrich, Uwe, Steinbeck, Christoph, Koepler, Oliver, Stegle, Oliver, Weimann, Joachim; Schörner-Sadenius, Thomas; Gutt, Christian; Stahl, Florian; Wagemann, Kurt; Schrade, Torsten; Schmitt, Robert; Eberl, Chris; Gauterin, Frank; Schultz, Martin; Bernard, Lars. (2019). Berlin Declaration on NFDI Cross-Cutting Topics (Version 1). Zenodo. <https://doi.org/10.5281/zenodo.3457213>
5. Nationale Forschungsdaten Infrastruktur (NFDI) - Sections. <https://www.nfdi.de/sections/?lang=en> (2023-04-24)
6. Barker, M., Wilkinson, R. & Treloar, A. The Australian Research Data Commons. *Data science journal* 18 (2019). <https://doi.org/10.5334/dsj-2019-044>
7. NCDI Cancer Research Data Commons. <https://datascience.cancer.gov/data-commons> (2023-04-24)
8. European Open Science Cloud (EOSC). <https://digital-strategy.ec.europa.eu/en/policies/open-science-cloud> (2023-04-24)
9. Global Open Research Commons IG. <https://www.rd-alliance.org/groups/global-open-research-commons-ig> (2023-04-24)

10. European Commission, Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K., et al., EOSC interoperability framework : report from the EOSC Executive Board Working Groups FAIR and Architecture, Publications Office, 2021, <https://doi.org/10.2777/620649>
11. EOSC Declaration (2017). https://eosc-portal.eu/sites/default/files/eosc_declaration.pdf (2023-04-24)
12. International Data Spaces Association (IDSA). <https://internationaldataspaces.org/> (2023-04-24)
13. FAIR Data Spaces. <https://www.nfdi.de/fair-data-spaces/?lang=en> (2023-04-24)
14. Basic Services for NFDI. <https://base4nfdi.de/> (2023-04-24)
15. Nationale Forschungsdaten Infrastruktur (NFDI) - Section Common Infrastructures. <https://www.nfdi.de/section-infra/?lang=en> (2023-04-24)
16. NFDI4Biodiversity. <https://www.nfdi4biodiversity.org/de/> (2023-04-24)
17. NFDI4Microbiota. <https://nfdi4microbiota.de/> (2023-04-24)
18. NFDI4Datascience. <https://www.nfdi4datascience.de/> (2023-04-24)
19. NFDI4Earth. <https://www.nfdi4earth.de/> (2023-04-24)
20. ARUNA. <https://aruna-storage.org> (2023-04-25)
21. Dehghani Z (2020) Data Mesh Principles and Logical Architecture - <https://martinfowler.com/articles/data-mesh-principles.html> (2023-04-24)

Galaxy and RDM

Being more than a workflow manager: living the data life cycle

Sebastian Schaaf^{1,2}[\[https://orcid.org/0000-0003-2982-388X\]](https://orcid.org/0000-0003-2982-388X), Anika Erxleben-Eggenhofer¹[\[https://orcid.org/0000-0002-7427-6478\]](https://orcid.org/0000-0002-7427-6478), and Björn Grüning¹[\[https://orcid.org/0000-0002-3079-6586\]](https://orcid.org/0000-0002-3079-6586)

¹ Freiburg Galaxy Team, Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany

² ELIXIR Officers Team, Institute for Bio- and Geosciences 5, Research Center Jülich, Germany

Abstract. The increasing amount of data generated by scientific research poses the challenge of providing an adequate infrastructure and tools that facilitate FAIR (Findable, Accessible, Interoperable and Reusable) data access, manipulation, analysis and visualization. Often, the burden of managing the metadata associated with the original data and the analysis lies with the researchers.

The open source Galaxy platform [1] is well-known for supplying tools and workflows for reproducible and transparent data analysis across scientific disciplines. It is a multi-user environment which facilitates sharing of e.g. tools, workflows, notebooks, visualizations, and data with others. There are three large Galaxy instances (US, Europe [2] and Australia) used by hundreds of thousands of researchers worldwide and that are using PBs of data. Galaxy handles the metadata transparently, releasing scientists from the burden and making it less prone to human errors. These features can be used without technical background by using a web browser or, for experts, through the Galaxy API.

Galaxy is more than a workflow manager: it provides scientists with access to reference data, databases (ENA, UniProt, NCBI, PDB, Ensembl...), external repositories (FTP, SFTP, Dropbox...), data sources through standard APIs (TRS, DRS from GA4GH). Importantly, Galaxy not only uses the metadata of input data, but enriches the metadata space with analysis information. Thus, users are finally enabled to apply export mechanisms for the data, targeting external resources (S3, ENA...). Beyond, also the applied workflow invocation can be exported invoking standardized formats (RO-Crate, BioComputeObject ...; [3]), effectively easing the sharing of the research artifacts, but also the details on the underlying analysis.

Although easy and standardized import/export functionalities are crucial features, entire data life cycles can be mapped into Galaxy. From the users' perspective, the built-in sharing features for data, workflows, histories etc. appear much more relevant in daily research efforts, making it particularly easy to reproduce results in order to verify their correctness and enable other researchers to build upon them in future studies. In fact, one of Galaxy's key features is its emphasis on **transparency, reproducibility, and reusability**. All provenance information of a dataset, including versions of used tools, parameters, execution environment, is captured and can be reused or exported using standards like BCO or RO-Crate to public archives. This highest level of provenance tracking also enables **traceability** and can also be used to reduce the environmental impact of a data analysis.

In addition to reproducibility, the Galaxy project places a strong emphasis on research data management (**RDM**; [4]). Beyond platform tools for data import, organization, sharing,

annotation, and export, data can be stored and accessed through a variety of providers, including cloud storages like NextCloud. This allows researchers to work with large datasets without the need for local storage infrastructures. The project encourages researchers to share their data and analysis workflows with the wider scientific community, with the aim of accelerating scientific discovery and innovation by following the FAIR principles. Notably, public instances offer not only computing capacities to users, but also persistent disk space, decoupling researchers from dependencies on local capacities and technical challenges (resources, capacities, support, ...). This enables democratizing data analysis in large. In practice, centralized and efficient user support is an important aspect of intuitive and borderless sharing in data science. Failures in tools or analysis procedures can be fixed for numerous users simultaneously and users can be nudged to use more efficient or correct tools.

In terms of RDM in Galaxy we will put a focus on **RO-Crate** (Research Object Crate; [5]) as a relatively recent development, implementing to a practical extent the FAIR Digital Objects (FDO) concept. RO-Crate enables researchers to organize and package their research data and other digital resources in a way that makes it easier to share, reuse, and reproduce their work; obviously this shows a great overlap with Galaxy's principles. On the European level, ELIXIR mandates the increased usage of RO-Crate. Notably, also in the German NFDI space decisions have been made for pushing RO-Crate, with DataPLANT being an early adopter. A central mission of DataPLANT, the first-round NFDI-funded consortium for connecting plant researchers in Germany, is to provide a suitable infrastructure for data analysis.

Galaxy has been widely adopted by the research community, particularly in fields such as data science, bioinformatics, and digital humanities. It has also been recognized as a key pillar of various European organizations such as the European Open Science Cloud (EOSC). The Freiburg Galaxy Team, being a central pillar for operating the European Galaxy server 'UseGalaxy.eu', is part of these efforts and also involved in further NFDI consortia. The European Galaxy server is a flagship project of the German Network of Bioinformatics infrastructure (de.NBI) and part of multiple CRCs. Since 2022 the European Galaxy community is officially part of EOSC with its own project 'EuroScienceGateway' [6], where OpenAIRE, the EOSC organization behind Zenodo, is an important partner. They bring in their knowledge graph, aggregating research data properties (metadata, links), supporting open science principles.

Our talk will describe how the Galaxy platform assists researchers from diverse technical backgrounds and scientific domains throughout the entire data life cycle: data access, processing, analysis, preservation, sharing and re-use (Figure 1). We will highlight cross-connections with other popular services and sketch future directions.

Keywords: RDM, Galaxy, data life cycle, reproducibility, RO-crate

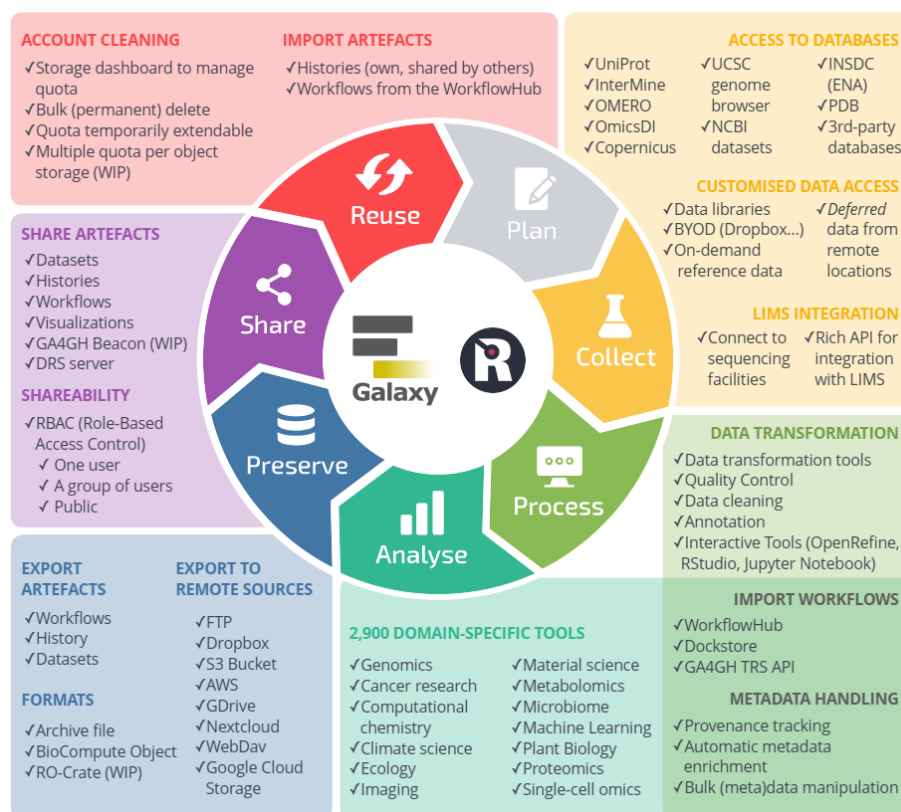


Figure 1. Data life cycle invoking Galaxy and RO-crate.

Competing interests

The authors declare that they have no competing interests.

References

1. The Galaxy Community - Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update; Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351, <https://doi.org/10.1093/nar/gkac247>
2. "Galaxy | Europe." <https://usegalaxy.eu/> (accessed Apr. 24, 2023).
3. <https://galaxyproject.org/news/2023-02-23-structured-data-exports-ro-bco/> (accessed Apr. 24, 2023).
4. https://rdmkit.elixir-europe.org/galaxy_assembly (accessed Apr. 24, 2023).
5. S. Soyland-Reyes et al. "Packaging research artefacts with RO-Crate" Data Science, vol. 5, no. 2, pp. 97-138, 2022, <http://dx.doi.org/10.3233/DS-210053>
6. <https://galaxyproject.org/projects/esg/> (accessed Apr. 24, 2023).

RADAR: Building a FAIR and Community Tailored Research Data Repository

Felix Bach¹, Kerstin Soltau¹, Sandra Göller¹, Christian Bonatto Minella¹ and Stefan Hofmann¹

¹ e-Research Institute, FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany

Abstract. The research data repository [RADAR](#) is designed to support the secure management, archiving, publication and dissemination of digital research data from completed scientific studies and projects. Developed as a collaborative [project funded](#) by the [Deutsche Forschungsgemeinschaft](#) (DFG, German Research Foundation) (2013-2016), the system is operated by [FIZ Karlsruhe](#) - Leibniz Institute for Information Infrastructure - and currently serves as a generic cloud service for about 20 universities and non-university research institutions. Since its launch, RADAR has witnessed significant changes in the landscape of research data repositories and the evolving needs of researchers, research communities and institutions. In our presentation within the “Enabling RDM” Track, we will show how RADAR is responding to these dynamic changes. In order to create a sufficiently large user base for the sustainable operation of the system, we have moved RADAR away from its previous single focus on a discipline-agnostic cloud service and towards a demand-driven functional optimisation. In 2021, we introduced an additional operating model for institutions ([RADAR Local](#)), where we operate a separate RADAR instance locally at the institution site exclusively using the institutional IT-infrastructure. In 2022 we opened up RADAR to new target groups with community-specific service offerings, in particular in the context of the [National Research Data Infrastructure](#) (NFDI). Beside the expansion of the functional scope, our ongoing development work focuses also on strengthening the system's support for the FAIR principles [1] and the concepts of [FAIR Digital Objects](#) (FDO) [2] and [Schema.org](#). Our presentation will outline recent RADAR developments and achievements as well as future plans thus providing solutions and synergy potential for the scientific community and for other service providers.

Keywords: NFDI, Chemistry, Culture, Infrastructure

1. Enhancements of the RADAR metadata schema

Aiming to enrich the descriptive metadata and strengthen the implementation of the FAIR principles, two important standard data - the [Research Organisation Registry](#) (ROR) and the “[Gemeinsame Normdatei](#)” (GND, Integrated Authority File) - were implemented into the generic [RADAR metadata schema](#) in December 2023. Besides, further controlled vocabularies (e. g. for licence information and related resources) were introduced. These enhancements promote in particular findability and interoperability of published datasets. For the same reason, the integration of the [TIB Terminology Service](#), which helps encoding the [DFG Classification of Subject Areas Ontology](#) (DFGFO) into an ontology, is currently under consideration as a useful standardisation measure for a future update of the RADAR schema. Moreover, we intend to explore different FAIR assessment tools to optimise and finetune our future developments.

2. Subject-specific metadata schemas

In 2022, we launched our community specific offerings [RADAR4Chem](#) and [RADAR4Culture](#) to provide researchers in the fields of chemistry ([NFDI4Chem](#)) and cultural studies ([NFDI4Culture](#)) with free data publication services. Similar sustainable research data publishing offerings in other scientific communities are underway.

It is especially in these subject-specific contexts that the established generic RADAR metadata schema needs to be opened up. Currently, this requirement is already supported by an upload option for subject-specific metadata files (XML) that are validated against configurable schemas given as XSD files.

However, we aim to enable those metadata annotations via more user friendly input masks which is currently ongoing work. For this purpose, we are extending the RADAR backend to support JSON based metadata (i.e. JSON-LD) in addition to XML, which will give users greater flexibility in creating and/or importing their own discipline-specific metadata in manifold schemas into RADAR. Nevertheless, a great challenge for all subject-based repositories in this context is to agree upon standards in the rapidly changing landscape and to make the existence of subject-specific metadata visible in evidence systems for publications such as DataCite that are focused on generic metadata.

3. Embedded metadata on dataset landing page

In addition to submitting descriptive metadata of published datasets to DataCite, we offer metadata harvesting via our [OAI provider](#) for all RADAR data publications, including discipline-specific metadata, according to the standardised [Open Archives Initiative Protocol for Metadata Harvesting](#) (OAI-PMH). This supports the dissemination, visibility and discoverability of RADAR research data and enables research communities to integrate RADAR content into specialist portals or knowledge graphs. However, findability, interoperability and in particular machine-readability are further maximised by applying established standards such as Schema.org. Moreover, new alternative approaches such as data visiting, Fair Digital Objects (FDO), and [FAIR signposting](#) are gaining significance as promising ways to increase machine-actionability, e.g. to allow distributed machine learning.

Data visiting targets landing pages directly, allowing machines to dynamically access both metadata and data. FAIR signposting takes a low-threshold approach that allows machines to interact with scholarly portals in a unified way, bypassing proprietary APIs and following typed links (e. g. DOI, author PIDs, web addresses of landing pages and descriptive metadata) in the [HTTP link headers](#). Late 2022, we enriched the source code of all dataset landing pages in RADAR, improving their machine readability and -actionability.

In the following software releases in early 2023, we optimised our embedded metadata into the landing pages for datasets using Schema.org annotation in JSON-LD and in Turtle format relevant for facilitating search, integration, and analysis of metadata of research datasets. In the future, we also plan to evaluate the [RO-Crate](#) approach (a community effort to practically achieve FAIR packaging of research objects) with their structured metadata to optimise our current preservation container solution based on [BagIt-standard](#).

4. Summary

In summary, the RADAR data repository has evolved to meet the changing needs of researchers and institutions, with a focus on demand-driven functional optimisation, community-specific service provision and increased support for concepts such as the FAIR principles, Schema.org and FDO. Subject-specific metadata annotation and the exploration of alternative approaches in contrast to traditional metadata harvesting have been achieved, nevertheless developments

continue. Future work includes the integration of further community-specific terminologies and ontologies as well as the exploration of FAIR assessment tools to improve research data published in RADAR.

Overall, RADAR aims to provide a complete solution for the management, archiving and publishing of research data for the scientific community and supporting best practises for other service providers.

Data availability statement

-

Underlying and related material

-

Author contributions

-

Competing interests

Authors declare no competing interests.

Funding

-

Acknowledgement

We would like to extend our gratitude to NFDI4Chem (DFG) – 441958208 and NFDI4Culture (DFG) – 441958017 for their financial support.

References

1. Wilkinson M., Dumontier M., Aalbersberg I., The FAIR Guiding Principles for scientific data management and stewardship (2016). *Sci Data* 3: 160018. <https://doi.org/10.1038/sdata.2016.18>
2. Schultes, E., Wittenburg, P. FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. In: Manolopoulos, Y and Stupnikov, S (eds.), *Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2018*. (2018) *Communications in Computer and Information Science*, 1003. Cham: Springer. DOI: <https://doi.org/10.1007/978-3-030-23584-0>

Toward the development of NII RDC application profile using ontology technology

Yasuyuki Minamiyama¹[\[https://orcid.org/0000-0002-7280-3342\]](https://orcid.org/0000-0002-7280-3342), Masaharu Hayashi¹[\[https://orcid.org/0000-0002-1451-9300\]](https://orcid.org/0000-0002-1451-9300), Ikki Fujiwara¹[\[https://orcid.org/0000-0001-6841-5243\]](https://orcid.org/0000-0001-6841-5243), Jun-ichi Onami¹[\[https://orcid.org/0000-0003-0790-8313\]](https://orcid.org/0000-0003-0790-8313), Shigetoshi Yokoyama¹[\[https://orcid.org/0000-0003-0060-8166\]](https://orcid.org/0000-0003-0060-8166), Yusuke Komiyama¹[\[https://orcid.org/0000-0001-6468-3718\]](https://orcid.org/0000-0001-6468-3718), and Kazutsuna Yamaji¹[\[https://orcid.org/0000-0001-6108-9385\]](https://orcid.org/0000-0001-6108-9385)

¹ National Institute of Informatics, Japan

Abstract. In recent years, there has been international progress in developing platforms that support the reproducibility and reusability of research data. Typical platforms adopt a service architecture integrating multiple information systems to cover the entire research data lifecycle. In realizing this architecture, specifications for inheriting processes and results executed on different information systems play an essential role. This study introduces our practices for application profile development using ontology technology in the NII Research Data Cloud.

Keywords: Research Data Platform, Application Profile, Interoperability, Ontology

1. Introduction

In recent years, there has been international progress in developing platforms that support the reproducibility and reusability of research data. For example, EUDAT adopts a service architecture integrating multiple information systems to cover the entire research data lifecycle [1]. Even in Japan, the NII Research Data Cloud (NII RDC) has been under development since 2017, led by the National Institute of Informatics [2].

In realizing this architecture, ensuring data interoperability is becoming the next challenge. If the meaning and structure assigned to research data generated by one information system cannot be adequately inherited, a different information system cannot interpret the data. This failure causes a severe loss of reproducibility and reusability of the research.

To address this issue, developing an application profile is essential in handling the meaning and structure of data between different information systems [3]. In RDM platform development, research data interoperability could be maximized by developing application profiles. However, there is no established procedure for constructing application profiles. This study introduces the process for developing an application profile using ontology theory and technology through our NII RDC development experience.

2. Approach

This chapter discusses the application profile development in line with ontology theory. We adopted the seven steps proposed by Noy & McGuinness as a development process [4]. We introduce “Step 3: Enumerate important terms in the ontology” and “Step 4: Define the classes and the class hierarchy” as distinctive steps in this paper.

2.1 Step 3: Enumerate important terms in the ontology

This step extracts the key terms to be addressed in the application profile. As a premise, we use "user story" method for determining specific functional requirements for NII RDC. A user story is a description of a function that is valuable to the user of the system or software, as expressed in the following manner:

"As [a user persona], I want [to perform this action] so that [I can accomplish this goal]"

These stories include essential information for understanding the meaning and structure assigned to the data generated by an information system. We extracted necessary terms from 138 NII RDC user stories by splitting them into subject/verb/object.

2.2 Define the classes and the class hierarchy

This step defines the classes and hierarchical relationships for this application profile. In designing the classes and hierarchical relationships, we extended the Activity Streams 2.0 framework. Activity Streams 2.0 enables the research data lifecycle to be viewed as a series of "RDM-related activities." Table 1 shows the lists of the defined classes.

Table 1. List of the defined classes.

Types	Class	Definition
Actor Types	Person	https://www.w3.org/ns/activitystreams#Person
	Institution	https://www.w3.org/ns/activitystreams#Organization
	Funding Agency	https://www.w3.org/ns/activitystreams#Organization
	Application	https://www.w3.org/ns/activitystreams#Application
	Service	https://www.w3.org/ns/activitystreams#Service
Activity Types	Activity	https://www.w3.org/ns/activitystreams#Activity ; https://purl.org/rdm/ontology
Object and Link Types	Resource	https://www.w3.org/TR/activitystreams-vocabulary/#dfn-object
	Repository	https://www.w3.org/TR/activitystreams-vocabulary/#Place
	Project	https://www.w3.org/ns/activitystreams#context
	DataManagementPlan	https://www.w3.org/ns/activitystreams#Document
	Collection	https://www.w3.org/ns/activitystreams#Collection
	Event	https://www.w3.org/ns/activitystreams#Event
	Access Rights Information	http://purl.org/dc/terms/RightsStatement
	Document	https://www.w3.org/ns/activitystreams#Document

We assigned "Actor Types" as the subject, "Activity Types" as the verb, and "Object and Link Types" as the target. Also, we defined 72 verbs extracted in Section 2.1 as an extension of the "Activity" class. The URIs and definitions of each vocabulary are available at <https://purl.org/rdm/ontology>. Note that "access rights" was found to be a term not covered by Activity Streams 2.0, so we adopted "Rights Statement" defined by DCMI Metadata Terms.

3. Implementation

We described the NII RDC user story in JSON format based on the application profile developed in Chapter 2. Figure 1 shows an example.

```

1  EXAMPLE XX
2
3  {
4    "@context": [
5      "https://www.w3.org/ns/activitystreams",
6      "https://purl.org/rdm/ontology"
7    ],
8    "summary": "As a Principal Investigator, I want to create an appropriate Data
9    Management Plan so that I can comply the requirement by Funding Agency.",
10   "type": "create",
11   "actor": {
12     "id": "https://orcid.org/0000-0002-7280-3342",
13     "type": "Person",
14     "name": "Yasuyuki Minamiyama",
15     "role": "ProjectLeader"
16   },
17   "object": {
18     "id": "https://doi.org/10.20736/12345678",
19     "type": "DataManagementPlan",
20     "name": "NII RDC project",
21     "published": "2023-03-05T00:00:00"
22   },
23   "instrument": {
24     "type": "Service",
25     "name": "GakuNin RDM"
26   }
27 }

```

Figure 1. Example description based on NII RDC application profile.

The structure of "type," "actor," "object," and "instrument" is in line with Activity Streams 2.0. The original user stories are described as "summary" so that the correspondence can be checked in case of later modification.

4. Conclusion

Implementation based on this application profile will facilitate communication between NII RDC systems and record the semantics contained in their activities. This approach can potentially improve data interoperability and be used as primary data for constructing a knowledge graph that represents RDM. We plan to explore the possibility of utilization through future analysis.

Underlying and related material

The application profile and the related document will be available at the following URL: <https://purl.org/rdm/ontology>.

Author contributions

Y.M. substantially contributed to the study conceptualization and the design of methodology. M.H., I.F., J.O., S.Y., and Y.K. contributed to the validation of the results. K.Y. supervised and directed the project. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. S. de Witt, D. Lecarpentier, M. van de Sanden, and J. Reetz. "EUDAT - A Pan-European Perspective on Data Management," in 2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), pp. 1–5, 2017.
2. Research Center for Open Science and Data Platform (RCOS), National Institute of Informatics. "Overview of the NII Research Data Cloud," <https://rcos.nii.ac.jp/en/service/>, (accessed 2023-4-6).
3. Rachel Heery, Manjula Patel. "Application profiles: mixing and matching metadata schemas," *Ariadne Issue 25*, (2000), (accessed 2023-4-6).
4. Noy, N.F., & McGuinness, D.L. "Ontology Development 101: A Guide to Creating Your First Ontology," 2001, https://protege.stanford.edu/publications/ontology_development/ontology101.pdf, (accessed 2023-4-6).

LabIMotion Electronic Lab Notebook as Research Data Management tool in Catalysis

Paolo Dolcet¹, Mariam L. Schulte^{1,2}, Florian Maurer¹, Nicole Jung³, Rinu Chacko¹, Olaf Deutschmann^{1,2}, and Jan-Dierk Grunwaldt^{1,2}

¹ Institute for Chemical Technology and Polymer Chemistry, Karlsruhe Institute of Technology, Germany

² Institute of Catalysis Research and Technology, Karlsruhe Institute of Technology, Germany

³ Institute of Organic Chemistry, Karlsruhe Institute of Technology, Germany

Abstract . In the field of heterogenous catalysis, Electronic Lab Notebooks (ELNs) are only rarely employed, due to complex data structure and different needs of the community with respect to the typical features provided by wide-spread ELNs. On the contrary, LabIMotion, an extension of the open-source ELN Chemotion, adapts to the characteristic complex workflows in heterogenous catalysis; these encompass catalyst synthesis, adaptation of devices and testing rigs, activity measurements, material characterization (possibly also *in situ/operando*, at large scale facilities) and are complemented by mathematical modelling and simulation. Direct links to metadata catalogues like SciCat (for synchrotron/neutron characterizations) and advanced research data management tools like Adacta (for improved traceability of catalytic data, experimental setups and related resources) are envisioned. The adaptability of LabIMotion in the catalysis field is presented via the topical examples of Cu-based catalysts for methanol synthesis and noble metal-based emission control catalysts.

Keywords: Electronic Lab Notebook, Catalysis, Metadata

Extended abstract

From both an academical and an industrial point of view, Research Data Management (RDM) tools are nowadays of prime importance for efficient usage and optimal reusability of research data, and Electronic Lab Notebooks (ELNs) play a key role in this, since they allow all data concerning an experiment to be stored electronically and be easily accessible.

In the (heterogenous) catalysis development environment, proper RDM practices coupled with big data science have the potential to be of great impact in the rational design of new and/or improved catalysts. Up-to-date, on the other hand, ELNs have found only limited permeation in the catalysis research field. This is due to the fact that the development of a catalyst follows a complex and intertwined workflow, leading to a rich data structure that is not well captured in currently available ELNs. For example, a typical workflow usually includes processes of catalyst synthesis and preparation, material characterization (which might include *in situ* characterization also at large scale facilities, like synchrotrons and neutron sources), activity measurements (also in this case possibly at large scale facilities, with data registered under *operando* conditions) coupled with adaptation of experimental reactor configuration, and finally complemented by mathematical modelling and numerical simulations. The currently available solutions are ill-suited to efficiently represent such a manifold system and no clear-cut solution is available.

We present here LabIMotion, an extension of the open-source ELN Chemotion [1] used in the NFDI4Chem community. This ELN extension, on the contrary, efficiently addresses the specific needs of the catalysis development environment. To achieve this goal, LabIMotion provides the possibility of creating detailed research plans and it is structured into defined functional areas, covering aspects such as catalyst synthesis, characterization and activity testing. For advanced characterizations performed at synchrotron and neutron facilities, LabIMotion will also provide linkage to the metadata catalogue SciCat [2], widely employed in the DAPHNE4NFDI consortium [3]. Devices management tools are also integrated in the platform, additionally offering a direct link to advanced tools like Adacta [4], supported by the NFDI4Cat project [5] and leading to improved traceability of all data and metadata related to kinetic performance studies, including therefore device-related resources. All these different areas are interlinked via a unique sample ID, and combined in a streamlined workflow that help the users in efficiently creating a traceable “digital twin” of their experimental catalyst environment.

Topical examples (e.g., Cu-based catalysts for methanol synthesis and noble metal-based emission control catalysts) are presented to guide through the typical data entry scenario (as shown in Figure 1) and to illustrate the capabilities and adaptability of LabIMotion, also with respect to logging, tracking and sharing of (meta)data.

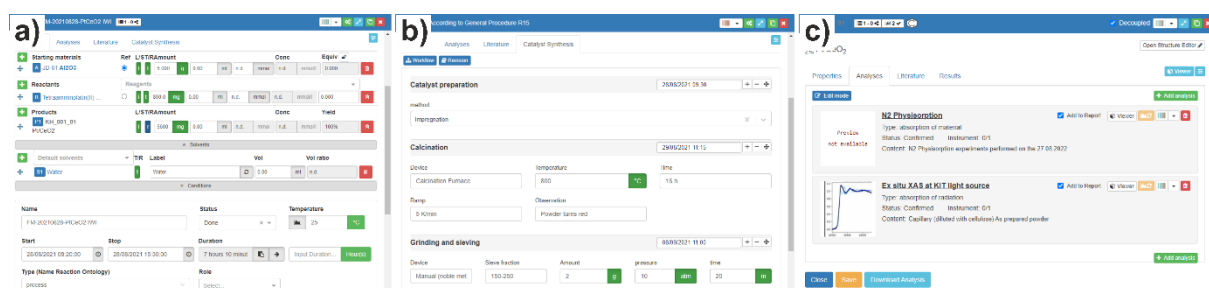


Figure 1. LabIMotion is organized in functional areas dedicated, e.g., to a) chemical information about the reactants, b) detailed synthesis procedure and c) material characterization.

LabIMotion is currently under advanced testing in the Collaborative Research Center 1441 “Tracking the Active Site in Heterogeneous Catalysis for Emission Control” [6] and in the Priority Programme 2080 “Catalysts and Reactors under Dynamic Conditions for Energy Storage and Conversion” [7], running at Karlsruhe Institute of Technology.

References

1. P. Tremouilhac, A. Nguyen, YC. Huang, et al. “Chemotion ELN: an Open Source electronic lab notebook for chemists in academia”. *J Cheminform* 9, 54 (2017), <https://doi.org/10.1186/s13321-017-0240-0>
2. <https://scicatproject.github.io/>
3. <https://www.daphne4nfdi.de>
4. H. Gossler, J. Riedel, E. Daymo, R. Chacko, S. Angeli, O. Deutschmann, “A New Approach to Research Data Management with a Focus on Traceability: Adacta”, *Chemie Ingenieur Technik*, 94, (2022), <https://doi.org/10.1002/cite.202200064>
5. <https://nfdi4cat.org/>
6. <https://www.trackact.kit.edu>
7. <https://www.spp2080.org>

Digitalization in Catalysis and Reaction Engineering: Automatizing Work Flows

Rinu Chacko¹, Hendrik Goßler², Johannes Riedel¹, Stephan Andreas Schunk³; Olaf Deutschmann¹

¹ Institute for Chemical Technology and Polymer Chemistry, Karlsruhe Institute of Technology, Germany

² omegadot Software & Consulting GmbH, Limburgerhof, Germany

³ hte GmbH, BASF SE, Germany

Keywords: Research Data Management, Reaction Engineering, Catalysis, Metadata

Extended abstract

Digitalization and surrounding efforts fostered by the advent of Industry 4.0 have been a development topic in the fields of chemistry and chemical engineering for several years – so timing may be right to look back and to question what the impact of efforts made, and to judge the impact of workflows and technologies developed. First and foremost, the first two key principle of Industry 4.0, namely *Interconnectivity* and *Information Transparency* play the crucial role in the context of Digitalization as only a seamless flow of data based on standardized formats remains key to enabling *Decentralized Decisions* and *Technical Assistance*. Digitization and digital tools play a key role in the acceleration data transfer and development efforts, the automation and autonomation of technical equipment employed, and the digital transformation of “classical” chemical reaction engineering processes towards an ideal originally projected by the high-tech agenda of Industry 4.0.

In this presentation we will start our journey with fully integrated environments on a laboratory level where data are not only made available in data warehouses but can be used to drive feedback loops to autonomously drive experimental devices. The concrete example shows how chemometrics, based on complex chemical analytics can be used to trigger autonomous decision making in order to harvest the maximum value of experimental resources [1]. We will continue with illustrative cases where we tap into the power of simulation software, that facilitates the understanding of physicochemical processes and assists in the optimization of design and operation points of chemical and electrochemical reactors. The development of the required kinetic models is an iterative process, which can be sped up by automatized work flows [2]. Furthermore, efficient research data management systems can ensure the (re)usability and traceability of experimental data for the development of novel catalysts and mechanisms [3]. To allow for in-depth insights into the underlying strategy of coupling of simulation and experiment, examples from the fields of reaction engineering including heterogeneous catalysis, electrochemistry and the production of alternative energy carriers are chosen.

A variety of tools have been developed and are under construction in the NFDI consortia such as NFDI4Cat [4]. Now, strong efforts are taken to combine the individual tools. Exemplarily, the Figure illustrates the communication between software tools that manage infor-

mation on catalyst synthesis, characterization, and performance as well as drivers for numerical simulation of catalytic reactors using these catalytic materials. Recently developed optimization tools to speed up model development and scale-up will soon extend this picture [5, 6].

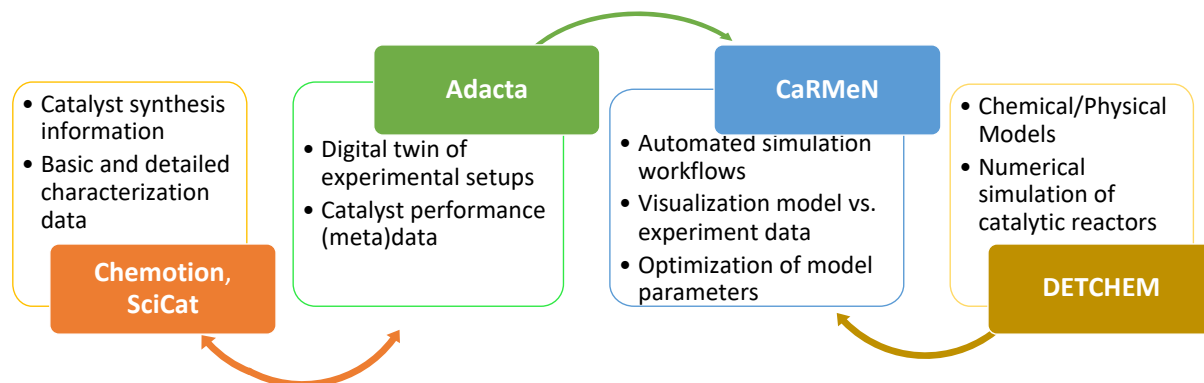


Figure 1: Functions and communication interfaces between software tools recently developed for digitalization in catalysis and reaction engineering.

The presentation is concluded by an overview of the work in the consortium NFDI4Cat, where digital tools, workflows and service offerings are developed in a community driven effort for catalysis related sciences including chemical engineering and process design.

Acknowledgment

This contribution including this abstract is based on a Plenary Talk at the Annual Meeting on Reaction Engineering 2023, Frankfurt, May 15-17, 2023. Parts of this work were supported by the NFDI4Cat project, which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with the project number 670389-NFDI 2/1.

References

- [1] M. Kirchmann, A. Haas, C. Hauber, S Vukojevic. PTQ Q4, 2015, 119-128.
- [2] H. Gossler, L. Maier, S. Angeli, S. Tischer, O. Deutschmann, PCCP, 2018, 20, 10857.
- [3] H. Gossler, J. Riedel, E. Daymo, R. Chacko, S. Angeli, O. Deutschmann, CIT, 2022, 94, 1798.
- [4] D. Linke, C. Wulf, M. Beller, T. Boenisch, O. Deutschmann, S. Hanf, N. Kockmann, R. Kraehnert, M. Oezaslan, S. Palkovits, S. Schimmler, S.A. Schunk, K. Wagemann. Chem-CatChem, 2021, 13, 3223.
- [5] R. Chacko, K. Keller, S. Tischer, A.B. Shirsath, P. Lott, S. Angeli, O. Deutschmann. J. Phys. Chem. C (2023), DOI: 10.1021/acs.jpcc.2c08179.
- [6] B. Kreitz, P. Lott, J. Bae, K. Blöndal, S. Angeli, Z.W. Ulissi, F. Studt, C.F. Goldsmith, O. Deutschmann. ACS Catal., 2022, 12, 11137.

Data Collections Explorer

An easy-to-use tool for sharing and discovering research data

Philipp Ost¹[\[https://orcid.org/0000-0002-7198-0566\]](https://orcid.org/0000-0002-7198-0566), Yusra Shakeel¹[\[https://orcid.org/0000-0001-5135-4325\]](https://orcid.org/0000-0001-5135-4325), and Philipp Tögel¹[\[https://orcid.org/0000-0002-1791-0268\]](https://orcid.org/0000-0002-1791-0268)

¹ Karlsruhe Institute of Technology, Germany

Abstract

There is a wide variety of archives, databases, and repositories currently available that provide access to research data. However, basic information about these systems is often difficult to gather, such as whether there are limits to the size of data sets that can be published or whether there is any publication fee that applies. In addition to that, there are plenty of research groups publishing their research data sets independently of these infrastructures, making it difficult for scientists to find them since they are not centrally registered. Research data must be easily discoverable and accessible for scientists to use it effectively. The Data Collections Explorer [1, 2], developed within the national research data infrastructure for the engineering sciences NFDI4Ing, is an easy-to-use information system addressing these needs. It is a low threshold information system that provides an overview of research data repositories, archives, databases as well as individually published data sets. Similar systems exist in other subject areas, for example the Data Repository Finder [3] focusing on the medical, life and social sciences. Contrary to the Data Collections Explorer, the Data Repository Finder only lists repositories.

By providing a low entrance barrier, the Data Collections Explorer makes it easy to share and discover repositories as well as data sets. Furthermore, scientists get a quick overview of the most important facts about services and data sets, such as access rights or usage restrictions. To further elaborate, we consider two practical use case scenarios:

1. Scientists searching for data sets: are there data sets available to aid in my research? Are there benchmarks available to check my results? Are these data sets available under an open access license? Are there usage or access restrictions?
2. Scientists aiming to publish data sets: among community-specific repositories, which ones are suitable to publish my research data? Do repositories restrict the size of the data sets that can be uploaded, and if so, what are the limits? Is there any publication fee charged and if so, how much is it?

To answer these questions, the Data Collections Explorer provides a free text search and filters for the type of service, subject area, and access license. Wherever appropriate and available, information on data size limits and publishing fees are given. Updates to the content are mostly initiated by the input of scientists, thus making sure the service fits their needs. Individual data sets, i.e., from the engineering sciences, are listed only if they are published outside the established infrastructure. Unlike re3data [4], services are listed irrespective of whether the operator is an individual, a community, or a legal entity. As a widely accepted authority [5], re3data is imposing strict criteria to list new entries. While smaller-scale projects,

such as databases or repositories run by individuals or communities, might still provide value to researchers, their lack of a legal entity makes them ineligible for listing in re3data [6]. In contrast to this, owing to the low-threshold approach, the Data Collections Explorer is thus capable to complement such established systems, as it provides coverage of entries beyond their scope, or which are not listed there.

Since its launch in March 2022, the Data Collections Explorer is being updated regularly based on the constructive feedback received from the engineering community, namely NFDI4Ing, but also NFDI4Chem and NFDI-MatWerk. After its launch, the Data Collections Explorer was presented in each NFDI4Ing Task Area. These task areas cover all aspects of engineering according to the classification by the German Research Foundation (DFG) [7]. Scientists provided feedback on their requirements of such a service, for instance on usability and access methods; contributions also included the repositories, databases, and data sets which are relevant to their work. Motivated by this positive feedback received during the past year, we are developing a new and improved version to overcome some limitations of the current approach. Currently, the Data Collections Explorer is a human-centered information system and not readily machine-accessible. To achieve the latter, we build a knowledge graph, c.f. Figure 1, that would not only allow easier machine-accessibility, but also a smoother integration with existing efforts within NFDI4Ing and the wider scientific community. Future work comprises improving the technical aspects, i.e., the development of user interfaces and APIs, but also maintenance and curation processes, i.e., to ensure the long-term sustainability of the service.

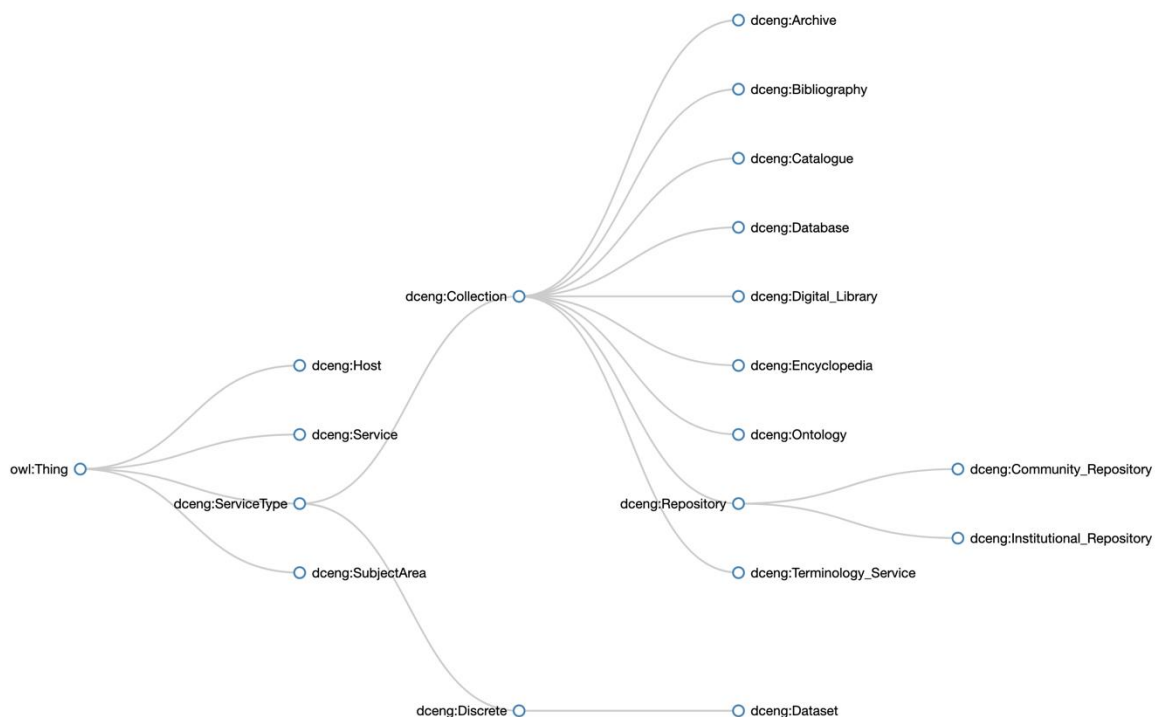


Figure 1 The new graph structure of the Data Collections Explorer.

Our proposed concept is not solely limited to the engineering sciences. The Data Collections Explorer has sparked attention in various NFDI consortia, as well as projects outside of NFDI. To broaden the impact of our concept, as a first step we are working on expanding the Data Collections Explorer to the materials science and engineering community within NFDI-MatWerk.

Competing interests

The authors declare that they have no competing interests.

Contributions

- Philipp Ost – Conceptualization, Investigation, Writing – original draft
- Yusra Shakeel – Writing – review & editing
- Philipp Tögel – Writing – review & editing

Funding

This work has been supported by NFDI4Ing (DFG – project number 442146713), NFDI-Mat-Werk (DFG – project number 460247524) and the Helmholtz Metadata Collaboration (HMC) platform.

References

- [1]. NFDI4Ing Data Collections Explorer, data-collections.nfdi4ing.de, last accessed: 2023-04-25
- [2]. Ost, P. (2022). Data Collections Explorer. Helmholtz Metadata Collaboration | Conference 2022, Online. <https://doi.org/10.5445/IR/1000151429>
- [3]. Data Repository Finder, data-repository-finder.ll.mit.edu, last accessed: 2023-07-17
- [4]. re3data.org - Registry of Research Data Repositories. <https://doi.org/10.17616/R3D> last accessed: 2023-04-25
- [5]. Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., Witt, M., Schirnbacher, P., Bertelmann, R., Scholze, F. (2017). The Landscape of Research Data Repositories in 2015: A re3data Analysis. D-Lib Magazine, 23(3/4). <https://doi.org/10.1045/march2017-kindling>
- [6]. Strecker, D., Bertelmann, R., Cousijn, H., Elger, K., Ferguson, L. M., Fichtmüller, D., Goebelbecker, H.-J., Kindling, M., Kloska, G., Nguyen, T. B., Pampel, H., Petras, V., Schabinger, R., Schnepf, E., Semrau, A., Trofimenko, M., Ulrich, R., Upmeier, A., Vierkant, P., Weisweiler, N. L., Wang, Y., Witt, M. (2021). Metadata Schema for the Description of Research Data Repositories: version 3.1, re3data, 37 p. <https://doi.org/10.48440/re3.010>
- [7]. The DFG subject area structure, https://www.dfg.de/en/research_funding/proposal_funding_process/interdisciplinarity/subject_area_structure/index.html, last accessed: 2023-07-18

The BAM Data Store

Piloting an openBIS-Based Research Data Infrastructure in Materials Science and Engineering

Rukeia El-Athman¹[\[https://orcid.org/0000-0003-0749-160X\]](https://orcid.org/0000-0003-0749-160X), Jörg Rädler¹[\[https://orcid.org/0000-0002-8935-6059\]](https://orcid.org/0000-0002-8935-6059), Oliver Löhmann¹[\[https://orcid.org/0000-0002-1527-1319\]](https://orcid.org/0000-0002-1527-1319), Angela Ariza¹[\[https://orcid.org/0000-0002-1005-5780\]](https://orcid.org/0000-0002-1005-5780) and Thilo Muth¹[\[https://orcid.org/0000-0001-8304-2684\]](https://orcid.org/0000-0001-8304-2684)

¹ Bundesanstalt für Materialforschung und -prüfung, Germany

Abstract. As a partner in several NFDI consortia, the Bundesanstalt für Materialforschung und -prüfung (BAM, German federal institute for materials science and testing) contributes to research data standardization efforts in various domains of materials science and engineering (MSE). To implement a central research data management (RDM) infrastructure that meets the requirements of MSE groups at BAM, we initiated the Data Store pilot project in 2021. The resulting infrastructure should enable researchers to digitally document research processes and store related data in a standardized and interoperable manner. As a software solution, we chose openBIS, an open-source framework that is increasingly being used for RDM in MSE communities.

The pilot project was conducted for one year with five research groups across different organizational units and MSE disciplines. The main results are presented for the use case “nanoPlattform”. The group registered experimental steps and linked associated instruments and chemicals in the Data Store to ensure full traceability of data related to the synthesis of ~400 nanomaterials. The system also supported researchers in implementing RDM practices in their workflows, e.g., by automating data import and documentation and by integrating infrastructure for data analysis.

Based on the promising results of the pilot phase, we will roll out the Data Store as the central RDM infrastructure of BAM starting in 2023. We further aim to develop openBIS plugins, metadata standards, and RDM workflows to contribute to the openBIS community and to foster RDM in MSE.

Keywords: Research Data Infrastructure, ELN, openBIS, Materials Science and Engineering.

1 Premise and Motivation

In the Bundesanstalt für Materialforschung und -prüfung (BAM, German federal institute for materials science and testing), research and development, as well as scientific and technical services are conducted in 10 departments divided into ~60 scientific units, comprising a scientific staff of more than 1,000 employees. As a partner in several NFDI consortia (currently in NFDI-MatWerk, FAIRmat, and DAPHNE4NFDI; prospective partner in NFDI4Chem), BAM contributes to research data standardization in various domains of materials science and engineering (MSE). Before introducing the Data Store, no software solutions for central research data management (RDM) or electronic lab notebooks (ELN) were established, resulting in institutional “data silos” that hampered the generation of FAIR datasets [1]. To fill this gap, we

launched the BAM Data Store project in 2021 to build an infrastructure that enables BAM researchers to document research processes and store associated data in a standardized and interoperable manner.

To collect RDM requirements and to test the implementation in the Data Store, we conducted a one-year pilot project with five research groups at BAM, consisting of 50 members from 15 scientific units. The pilot groups represent multidisciplinary within the MSE domain and routinely apply a broad range of methodologies.

2 openBIS: An Open-Source RDM and ELN Software

The Data Store is based on the open-source software openBIS [2], [3]. Developed by the Scientific IT Services of ETH Zurich, openBIS was originally intended as a framework to support RDM in life science laboratories but is increasingly being used in the MSE community. openBIS offers a web-based graphical user interface (GUI) for the digital representation of laboratory inventory, and an ELN for the documentation of experimental procedures. Data files of any format can be imported either via the GUI or via programming interfaces and linked to inventory elements and experimental steps. In addition, openBIS provides various interfaces for data transfer, e.g., for exporting meta(data) to the data repository Zenodo [4] as well as for analyzing data in the web-based environment Jupyter Notebook [5].

3 Pilot Project: Testing openBIS with Five MSE Research Groups

3.1 IT Infrastructure

openBIS is designed to run on Linux servers and requires a Java environment and a PostgreSQL database system. To gain maximum flexibility of the configuration, every pilot group was equipped with a dedicated virtual Linux server containing an openBIS installation managed by the central IT department. While all datasets are stored as regular files on an openBIS server (which may require a large amount of local storage), the data models, metadata, relations, and permissions are stored in the database. A JupyterHub installation was added to each server for testing purposes and configured to work seamlessly with openBIS. Central file shares and individual import scripts (a.k.a. Dropboxes) were used to automate the import and partial analysis of data from lab equipment.

3.2 Onboarding Process

The onboarding to the Data Store was conducted individually for each pilot group. In a first workshop, the group members defined and prioritized their RDM requirements as testable user stories along the phases of the data lifecycle. Next, the group's admins learned about the underlying data hierarchy of openBIS and were tasked with the creation of a data model to digitally represent their group's inventory and experimental steps. After the inventory was completed and the building blocks for the lab notebook were defined, the whole group was trained to use the Data Store for their daily work. The pilot phase was concluded by testing the user stories. The implementation of the Data Store by the pilot group "nanoPlattform" is described below.

3.3 Use Case "nanoPlattform"

The "nanoPlattform" is an internally funded BAM project with the aim to provide reference materials, reference data, and reference protocols in the field of nanoscience. Within the project, more than 30 scientists and technical staff of different disciplines from 10 different scientific units currently use the Data Store to ensure data traceability and data transfer. Within one year, about 400 nanomaterials were synthesized and more than 1000 experimental steps were

registered using more than 90 instruments and more than 150 different chemicals. All these objects are stored and connected in the Data Store, ensuring full traceability of the scientific process and full accessibility for all group members (Figure 1).

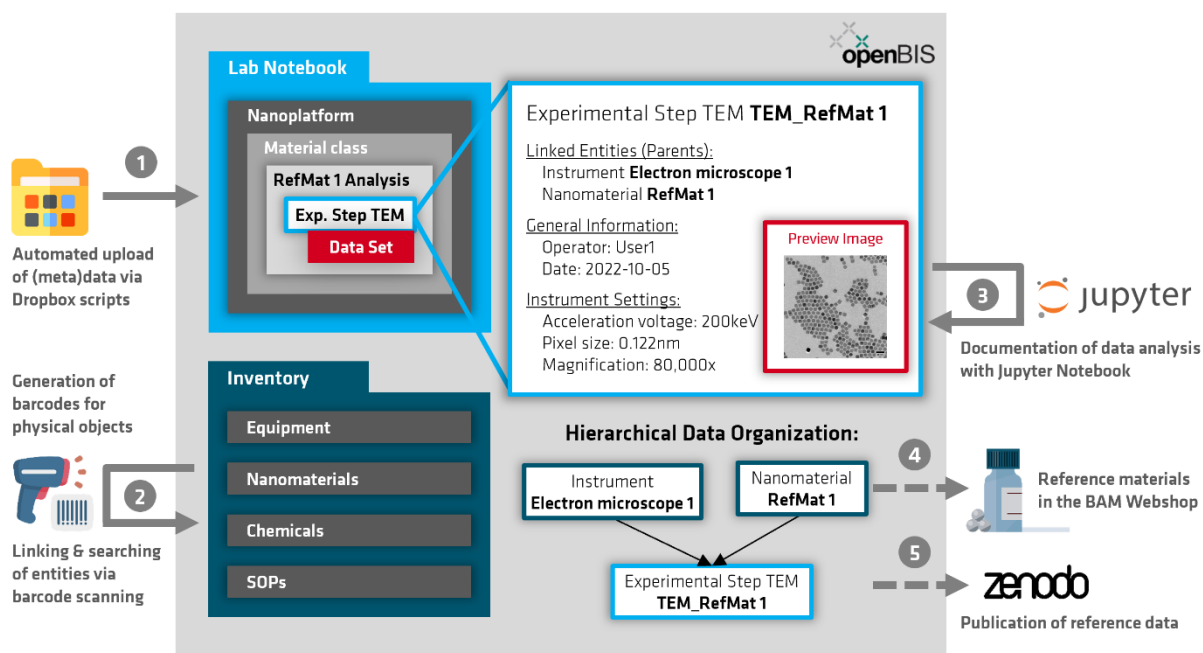


Figure 1. Data Store implementation of the pilot group “nanoPlatform”.

The synthesis of nanoparticles is documented in the lab notebook of the Data Store and linked to the chemicals registered in the inventory. (1) Measurement data are imported and attached to the respective experiment by Dropbox scripts which also parse and register the associated metadata. (2) Synthesized samples are labeled with barcodes generated by openBIS, allowing all project members to access sample-related metadata and data simply by scanning a barcode. Sample characterizations are also documented in the Data Store and linked to the respective samples, instruments, sample preparation. (3) Data are analyzed using Jupyter Notebooks that access data via the Python extension pyBIS. The resulting Notebooks are saved in the Data Store and linked to the respective experiments. (4) The synthesized reference material is available for purchase in the BAM Webshop. (5) The publication of reference data and protocols are currently in preparation. (Left-side icons made by Freepik from www.flaticon.com.)

4 Conclusion and Outlook

As illustrated by the example of the “nanoPlatform”, many of the RDM requirements of MSE research groups can be met by openBIS functionalities. The pilot project showed that the transition to a digital system for the representation of laboratory notebooks requires researchers to invest time and effort during the implementation and customization phase. However, the Data Store can ultimately save time by automatizing daily tasks, e.g., the barcode-supported documentation of consumables used for measurements, and by the automatic import of standardized (meta)data from instruments. In addition, the Data Store enables researchers to implement RDM practices according to the FAIR principles by integrating software tools such as Jupyter Notebook to document data analysis, associated code, and underlying experimental data.

Altogether, the Data Store facilitates cooperation between scientists and across research communities within BAM by providing a common infrastructure and shared metadata schemas.

Based on the successful implementation of openBIS in research workflows during the pilot project, the Data Store will be rolled out as the institute's central RDM infrastructure from 2023. We expect the Data Store to become the backbone of digital RDM at BAM and we aim to develop additional openBIS plugins, metadata standards, and RDM workflows that will benefit various MSE communities both within and beyond BAM.

Data availability statement

No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Author contributions

Rukeia El-Athman: Methodology (lead); conceptualization (equal); writing – original draft (lead); writing – review and editing (equal); visualization (lead). Jörg Rädler: Software (lead); conceptualization (equal); writing – original draft (supporting); writing – review and editing (equal). Oliver Löhmann: Methodology (supporting); writing – original draft (supporting); writing – review and editing (equal). Angela Ariza: Writing – original draft (supporting); writing – review and editing (equal). Thilo Muth: Supervision (lead); project administration (lead); conceptualization (equal); writing – review and editing (equal).

Competing interests

The authors declare that they have no competing interests.

Funding

The BAM Data Store project is funded by internal BAM resources.

Acknowledgement

We would like to thank Caterina Barillari from the Scientific IT Services of ETH Zurich for her support and advice, and for conducting the openBIS webinars during the pilot phase.

References

1. M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, no. 1, Art. no. 1, Mar. 2016, doi: 10.1038/sdata.2016.18.
2. A. Bauch et al., "openBIS: a flexible framework for managing and analyzing complex data in biology research," *BMC Bioinformatics*, vol. 12, no. 1, p. 468, Dec. 2011, doi: 10.1186/1471-2105-12-468.
3. C. Barillari, D. S. M. Ottoz, J. M. Fuentes-Serna, C. Ramakrishnan, B. Rinn, and F. Rudolf, "openBIS ELN-LIMS: an open-source database for academic laboratories," *Bioinformatics*, vol. 32, no. 4, pp. 638–640, Feb. 2016, doi: 10.1093/bioinformatics/btv606.
4. European Organization For Nuclear Research and OpenAIRE, "Zenodo." CERN, 2013. doi: 10.25495/7GXX-RD71.
5. M. Beg et al., "Using Jupyter for reproducible scientific workflows," *Comput. Sci. Eng.*, vol. 23, no. 2, pp. 36–46, Mar. 2021, doi: 10.1109/MCSE.2021.3052101.

6. "BAM-N012 - Nanomaterialien - Referenzmaterialien." https://webshop.bam.de/webshop_de/referenzmaterialien/nanomaterialien/bam-n012.html (accessed Apr. 18, 2023).

Current Insights from Task Area 1 in NFDI4Energy: Building and Serving the Energy Research Community

Oliver Werth¹[\[https://orcid.org/0000-0002-6767-5905\]](https://orcid.org/0000-0002-6767-5905), Stephan Ferenz²[\[https://orcid.org/0000-0001-9523-7227\]](https://orcid.org/0000-0001-9523-7227), Astrid Nieße²[\[https://orcid.org/0000-0003-1881-9172\]](https://orcid.org/0000-0003-1881-9172), Reinhard German³[\[https://orcid.org/0000-0002-9071-4802\]](https://orcid.org/0000-0002-9071-4802), Ludwig Hülk⁴, Christof Weinhardt⁵[\[https://orcid.org/0000-0002-7945-4077\]](https://orcid.org/0000-0002-7945-4077), and Berthold Vogel⁶

¹ OFFIS e.V. – Institute for Information Technology, Oldenburg, Germany

² Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

³ Friedrich-Alexander-Universität Erlangen-Nürnberg, Nürnberg, Germany

⁴ Reiner Lemoine Institut, Berlin, Germany

⁵ Karlsruher Institute of Technology, Karlsruhe, Germany

⁶ Soziologisches Forschungsinstitut Göttingen (SOFI) e.V., Göttingen, Germany

1. Introduction and Motivations

Energy system research has become increasingly reliant on modeling and simulation approaches. These endeavors are enabled by continuously improving tools and methods for developing, maintaining and sharing models and data. Knowledge of how to better conduct, share and archive one's research has become increasingly complex and hard to manage for individual researchers or single research groups. Identifying and including relevant scientists from energy research, social sciences, and further disciplines is sometimes difficult. Furthermore, a plethora of best practices and guidelines exist on how to prepare data, models and results in ways that make them easier to discover, verify and build upon. To present a sustainable, problem-solving, technical solution for the energy research community, NFDI4Energy develops in Task Area 1 (TA1) two services of the NFDI4Energy platform. Consequentially, the intention of this abstract within the disciplinary track "Engineering" is to provide an overview of the development process with a special focus on *Competence* and *Best Practices*. In addition, it discusses interconnections with other Task Areas as well as the chances and challenges that are associated with those connections. The academic audience, e.g., from the Information Systems and (Software) Engineering domain at CoRDI 2023, can observe and discuss our proposed procedure with other community members. Furthermore, we expect interested individuals to compare the proposed procedures with their own, which can lead to meaningful discussions and knowledge-sharing situations within the Engineering domain.

2. Task Area Objectives and Procedures

TA1 called "Building and Serving the Energy Research Community", is responsible for four **(1-4)** main TA objectives that a mixture of different procedures will achieve: First **(1)**, it collects and updates platform requirements by the energy modeling and simulation communities. For this first object, we follow a requirements-driven process that takes one of the envisioned user groups – the energy research community - into focus in the design process of the platform [1]. As a result, we will first develop an interview guideline within the consortia and identify and

motivate relevant stakeholders for interview participation. Semi-structured single- and group interviews will be performed to collect meaningful requirements for the development process by qualitative content analysis (e.g., Flick [2]). Integrating the community into this process motivates researchers to use our NFDI4Energy platform later and in the long term.

Second **(2)**, TA1 continuously monitors, adapts and improves offerings according to community feedback. A structured review of the strengths and weaknesses of various (research) platforms will identify best practices for user incentivization and user feedback. Feedback mechanisms will be developed based on the resulting review and platform-integrated usage statistic tracking, including quantitative and/or qualitative criteria. An evaluation will take place with the tracked usage statistics along the development process, expert interviews, and platform user surveys to improve the long-term usability of the infrastructure.

Third **(3)**, TA1 creates a platform for sharing best practices, community guidelines and access points to community services. Here, TA1 is responsible for two services that NFDI4Energy will provide: As a first service, *Best Practices* should curate and present the current best practices from the energy system research community. As a second service, *Competence* will guide (unexperienced) researchers within the community to find suitable contact persons, e.g., owners of uploaded datasets within the NFDI4Energy platform. This will be realized through a database of scientific institutes, their members, and relevant industrial partners. Ideally, this database will be searchable and help identify the right research and transfer partners for, e.g., upcoming research projects. In addition, we will provide a platform evaluation workshop to gauge acceptance and usage of the platform. Since the NFDI4Energy platform will live from the participation of its intended stakeholders, TA1 develops a quality assurance process for content submissions.

Last and fourth **(4)**, TA1 develops guidelines, materials and tutorials for best practices in energy system research. Potential tutorial topics will be identified through the extensive requirements analysis before. As a result, NFDI4Energy will produce tutorials and best practices for the energy community and deploy them on the platform. This will align with Open Science and the FAIR principles that advocate for transparency and accessibility of methods, data and results but lack clear definitions and proceedings [3]. Based on these recommendations NFDI4Energy will provide hands-on examples, e.g., research data and software. We see this as a mix of descriptive and instructional content presented in interactive formats, if applicable.

3. Discussions and Conclusions

Generally, interconnections of TA1 can be seen on the requirements, services, and content layer. Intentionally, TA1 has strong interconnections to TA2 ("Integrating Society and Policy in Energy Research") and TA3 ("Transparency and Involvement of the Energy-Related Industry"), which examine crucial requirements for the platform from the point of view of society and policy in energy research and energy-related industry. While the requirements of these stakeholders are also important for the development process, a continuous communication information exchange is inevitable for the overall development process. Also, we expect to interact closely with the team of TA4 ("FAIR Data for Energy System Research") as they work on ontology and metadata standards that should be practical and useful for researchers in the field of energy research. These ontology and metadata standards must be implemented into the platform appropriately. Since *Simulations* in interdisciplinary energy research are a key service provided, continuous communication and constant workflows are needed with TA5 ("Simulation in Interdisciplinary Energy Research"), too. The NFDI4Energy platform must be of high usability for its intended stakeholders. Therefore, concrete use cases will be developed in TA6 ("Use Cases for Community Services") that can serve as a foundation for developing content for *Best Practices*. The NFDI4Energy consortium is aware of potential risks within the

overall work and in TA1 in particular. To avoid lacking support and interest in the energy system research community, a detailed requirements analysis will be performed to reflect opinions and feelings appropriately. Community workshops and presentations of different prototypes will support promoting the provided services. Figure 1 depicts the interconnectedness of TA1 within NFDI4Energy described before:

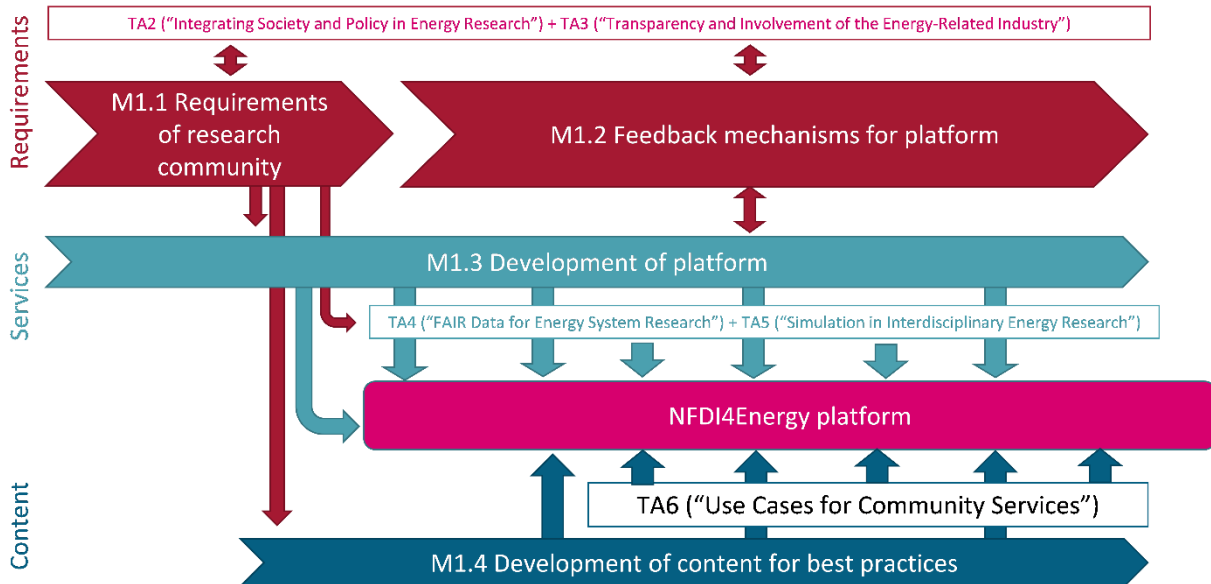


Figure 1. The interconnectedness of TA1 within the NFDI4Energy; Note: The four objectives of TA1 are internally referred to as “Measures” denoted here as M1.1 - M1.4.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number 501865131.

References

1. O. Werth, S. Ferenz, and A. Nieße, “Requirements for an open digital platform for interdisciplinary energy research and practice,” in *Proc. of the 15th International Conference on Wirtschaftsinformatik*, Nürnberg, 2022.
2. U. Flick, *An introduction to qualitative research*. London, UK: Sage, 2022.
3. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, et al., “The FAIR Guiding Principles for scientific data management and stewardship,” *Sci Data*, vol. 3, article 160018, Mar. 2016, doi: <https://doi.org/10.1038/sdata.2016.18>.

Building Ontologies and Knowledge Graphs for Mathematics and its Applications

Björn Schembera¹[\[https://orcid.org/0000-0003-2860-6621\]](https://orcid.org/0000-0003-2860-6621), Frank Wübbeling²[\[https://orcid.org/0000-0002-2375-2008\]](https://orcid.org/0000-0002-2375-2008), Thomas Koprucki³[\[https://orcid.org/0000-0001-6235-9412\]](https://orcid.org/0000-0001-6235-9412), Christine Biedinger⁴[\[https://orcid.org/0009-0002-5082-8386\]](https://orcid.org/0009-0002-5082-8386), Marco Reidelbach⁵[\[https://orcid.org/0000-0002-1919-1834\]](https://orcid.org/0000-0002-1919-1834), Burkhard Schmidt³[\[https://orcid.org/0000-0002-9658-499X\]](https://orcid.org/0000-0002-9658-499X), Dominik Göttsche¹[\[https://orcid.org/0000-0002-1552-497X\]](https://orcid.org/0000-0002-1552-497X), and Jochen Fiedler⁴[\[https://orcid.org/0000-0002-9176-780X\]](https://orcid.org/0000-0002-9176-780X)

¹Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, Stuttgart, Germany

²Institute of Applied Mathematics: Analysis and Numerics, University of Münster, Münster, Germany

³Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

⁴Fraunhofer Institute for Industrial Mathematics, Kaiserslautern, Germany

⁵Mathematics of Complex Systems, Zuse Institute Berlin, Berlin, Germany

Abstract: Ontologies and knowledge graphs for mathematical algorithms and models are presented, that have been developed by the Mathematical Research Data Initiative. This enables FAIR data handling in mathematics and the applied disciplines. Moreover, challenges of harmonization during the ontology development are discussed.

Keywords: Research Data Management, Mathematics, Ontologies, Knowledge Graphs

1 Introduction

Mathematical research data is vast, complex and multifaceted, and its typical appearances are formulae, models, data (numerical, symbolic, tabular), software and documents [1]. It emerges within the mathematical core sciences but also in the applied sciences, such as engineering, physics or even digital humanities. Given this, the Mathematical Research Data Initiative (MaRDI) is being established as the NFDI consortium of mathematics [2], [3]. Its mission is to develop a robust research data infrastructure for mathematics and beyond. It is here that MaRDI is setting up semantic technology (metadata, ontologies, knowledge graphs) and data infrastructures (portals, repositories) [4] for relevant information assets such as algorithms, models, problems, software or data to foster compliance with the FAIR principles [5].

MaRDI aims to support researchers solving real world problems by translating them into mathematical ones, applying adequate algorithms and transferring back the results (see figure 1). Here, several questions arise, e.g. existence of a mathematical model, availability of solving algorithms, input data or model validity. A non-negligible amount of time is required for finding existing models, algorithms and solvers.

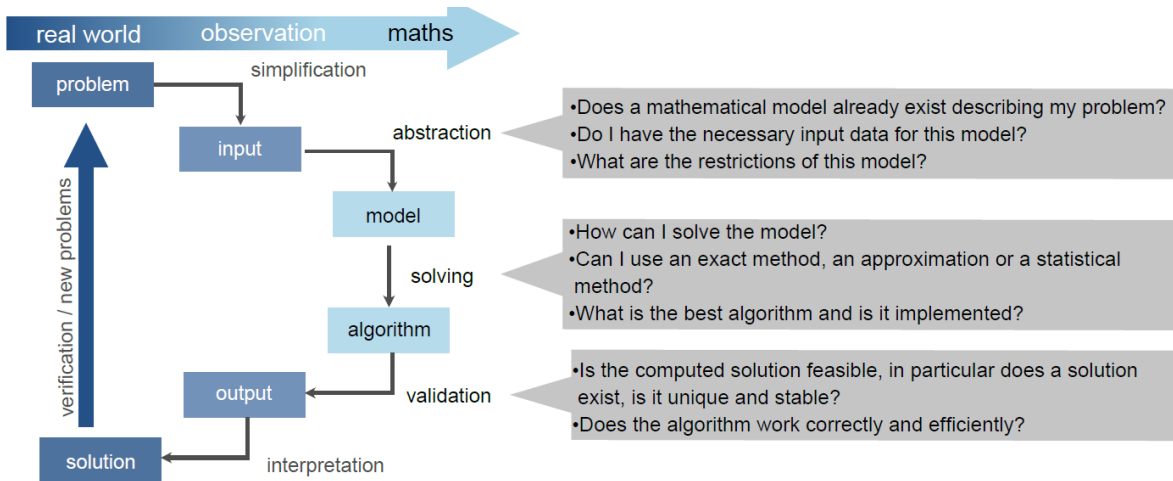


Figure 1. Typical modeling-simulation-optimization workflow considered in MaRDI and resulting competency questions [6].

2 Knowledge Graphs in Mathematics

Knowledge graphs are a well-established semantic technology for representing knowledge in a graph-structured data model, i.e. an ontology. Generally, they represent relations between objects as semantic triples, e.g. $(algorithm, solves, computable\ model)$, where *solves* is the relation describing that a specific algorithm acts as a solver for a computable model. Taking line planning in public transport as an example (cf. Lin-Tim [7]–[9]), different algorithms for optimization objectives like costs or passenger satisfaction are applicable. The integer program minimizing costs (the computable model), can then be solved by a branch-and-bound algorithm.

In the following, we present two ontologies for algorithms and models. These are linked to additional ontologies, e.g. for benchmarking and statistics.

2.1 A Knowledge Graph for Algorithms

Algorithms are a basic building block of applied mathematics. Algorithms solve problems, are implemented in software, and tested by benchmarks. Additionally, they are documented, invented, etc. in Publications.

The curated AlgoData knowledge graph [10] seeks to formalize the relations between the objects. It allows to answer questions of the form

- Which algorithms solve my problem?
- Which implementations are available? Is there a common analysis?
- Which benchmarks should I use for my problem?

The ontology has been kept as minimal as possible, with object classes algorithm, problem, software, benchmark and publication, and a total of 16 relations (see figure 2).

Initial graphs have been created in Control Configuration Selection, Model Order Reduction and Linear Algebra, for a total of 150 algorithms, documented in 825 publications. The current state is available on the AlgoData site [11]. The interface includes documentation of the ontology, keyword search, and a guided query tailored to the ontology. Technically, the frontend is based on a Apache Jena Fuseki with OWL reasoner and SPARQL interface to a Django server. A SPARQL endpoint can be provided.

Currently, the fundamental algorithms, problems etc. are provided by experts in the field. We envision editorial boards for each subdiscipline that keep this data up to date by accepting (or denying) proposals, made through the MaRDI portal. We have given major attention to ensuring that the processes for editors and users cause as little work as possible. A preview of the editorial process is available on the beta web site [12].

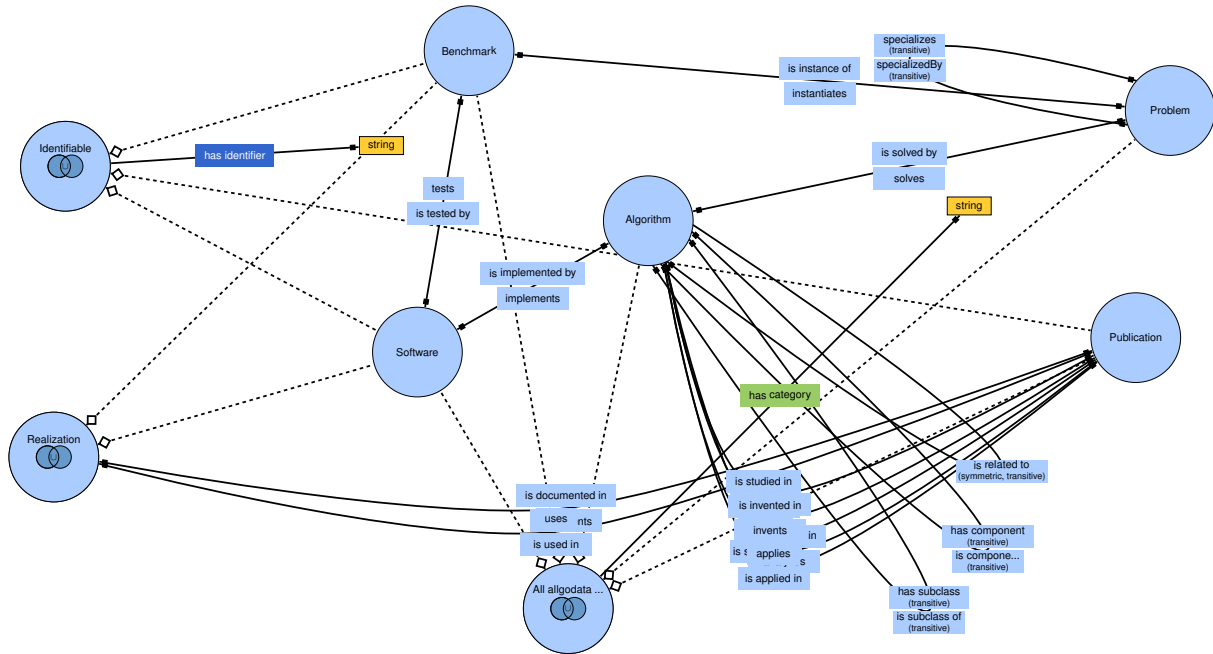


Figure 2. AlgoData ontology, visualized in VOWL

2.2 A Knowledge Graph for Models

Models are at the intersection of mathematics and multitude of other disciplines, serving as vital tools for understanding, predicting, and analyzing complex phenomena across various domains. Their importance stems from their ability to simplify real-world systems and represent them using mathematical equations which makes them computable. Because of their universal nature, models within different domains share common properties, such that developing or refining a model in one field can often lead to advancements and insights in others. By identifying these shared properties, researchers can bridge gaps between disciplines, fostering interdisciplinary collaboration and promoting the growth of knowledge in multiple areas simultaneously. Within MaRDI, the development of an ontology for mathematical models is advanced along case studies from manifold fields of the applied sciences, such as engineering or materials science. Through analysis of the case studies and their according workflows, it is possible to carve out relations, interdependencies and details of the underlying mathematical models. These are displayed on the MaRDI portal as a Wiki [13]. As of April 2023, the preliminary MaRDI model ontology includes the 8 classes (*model*, *computable model*, *application problem*, *application domain*, *equation*, *law*, *quantity*, *term*) and is shown in figure 3. The ontology can be seen as an extension of Model Pathway Diagrams [14], which are based on quantities (represented by terms) which are connected by laws (represented by equations). The ontology has been developed along the case studies, and first models, such as Navier-Stokes or diffusion from the mechanics domain, or Bayes from the stochastics domain have been integrated for testing purposes.

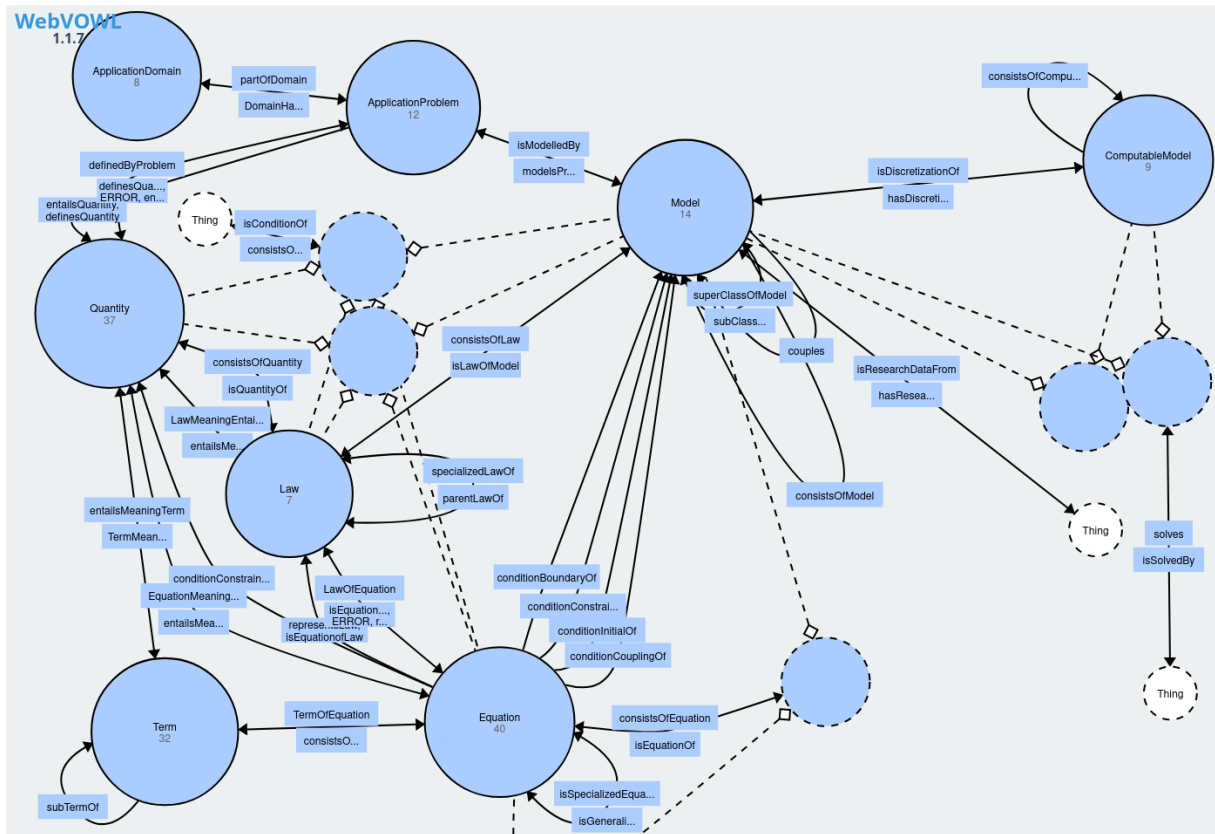


Figure 3. MaRDI Model Ontology classes and their relations. Visualisation done with WebVOWL.

3 Conclusion

In many areas of science, novel findings are acquired by processing and analyzing data. In general, these steps can be regarded as data transformations. Our mission in MaRDI is to structure and classify the possible mathematical data transformations and make them findable and accessible within mathematics and beyond.

To achieve this, we have presented ontologies and knowledge graphs for mathematical algorithms and models. In the future we will also integrate statistical algorithms and machine learning models. However, harmonization is one of the biggest challenges, not only between, but also within disciplines. As an example, the notion of a seemingly simple term like "problems" is controversial: in the models ontology, it represents the application problem, whereas in AlgoData, it stands for the pure mathematical problem. This could be solved by more precise labels and definitions. We see that especially across disciplines, harmonization requires ample coordination efforts, which can be provided, for example, through the NFDI.

Declarations

Data availability statement

The data in the knowledge graph for algorithms can be accessed via algodata.mardi4nfdi.de/. The data in the knowledge graph for models is as of now just for testing and will be published later.

Competing interests

The authors declare that they have no competing interests.

Funding

All authors are supported by MaRDI, funded by the Deutsche Forschungsgemeinschaft (DFG), project number 460135501, NFDI 29/1 "MaRDI – Mathematische Forschungsdateninitiative".

References

- [1] T. Boege, R. Fritze, C. Görgen, *et al.*, "Data Management Planning in the German Mathematical Community," *arXiv preprint arXiv:2211.12071*, 2022.
- [2] C. Görgen and R. Sinn, "Mathematik in der Nationalen Forschungsdateninfrastruktur," *Mitteilungen der Deutschen Mathematiker-Vereinigung*, vol. 29, no. 3, pp. 122–123, 2021.
- [3] P. Benner, M. Burger, D. Göddeke, *et al.*, "Die mathematische Forschungsdateninitiative in der NFDI: MaRDI (Mathematical Research Data Initiative)," *GAMM Rundbrief*, no. 1, pp. 40–43, 2022.
- [4] M. T. Horsch, S. Chiacchiera, W. L. Cavalcanti, and B. Schembera, *Data Technology in Materials Modelling*. Springer Nature, 2021.
- [5] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [6] The MaRDI consortium, *MaRDI: Mathematical Research Data Initiative Proposal*, May 2022. DOI: [10.5281/zenodo.6552436](https://doi.org/10.5281/zenodo.6552436). [Online]. Available: <https://doi.org/10.5281/zenodo.6552436>.
- [7] A. Schiewe, S. Albert, P. Schiewe, A. Schöbel, F. Spühler, and M. Stinzendörfer, *Documentation for LinTim 2022.08*, coursematerial, 2022. [Online]. Available: <http://nbn-resolving.de/urn:nbn:de:hbz:386-kluedo-69236>.
- [8] A. Schiewe, S. Albert, V. Grafe, P. Schiewe, A. Schöbel, and F. Spühler, *LinTim - integrated optimization in public transportation*. [Online]. Available: <https://www.lintim.net>.
- [9] M. Goerigk, S. Michael, and A. Schöbel, "Evaluating line concepts using travel times and robustness: Simulations with the LinTim toolbox," *Public Transport*, vol. 5, pp. 267–284, 2013. DOI: <https://doi.org/10.1007/s12469-013-0072-x>.
- [10] C. Himpe, H. Kleikamp, R. Fritze, and S. Rave, *MaRDI Task Area 2 - Scientific Computing @ WWU Münster. AlgoData - Algorithm Knowledge Graph - Ontology (Version 0.1)*, 2022. [Online]. Available: <https://mardi4nfdi.de/algoata/0.1>.
- [11] The MaRDI consortium, *AlgoData*, <https://algotata.mardi4nfdi.de/>, [Online; accessed 24-April-2023], 2023.
- [12] The MaRDI consortium, *AlgoData Editorial Access Beta*, <https://beta.m1.mardi.ovh/>, [Online; accessed 24-April-2023], 2023.
- [13] The MaRDI consortium, *MaRDI TA4 Portal*, <https://portal.mardi4nfdi.de/wiki/Portal/TA4>, [Online; accessed 24-April-2023], 2023.
- [14] T. Koprucki, M. Kohlhasse, K. Tabelow, D. Müller, and F. Rabe, "Model pathway diagrams for the representation of mathematical models," *Optical and Quantum Electronics*, vol. 50, pp. 1–9, 2018.

Modelling Scientific Processes with the m4i Ontology

Dorothea Iglezakis¹[\[https://orcid.org/0000-0002-8524-0569\]](https://orcid.org/0000-0002-8524-0569), Džulia Terzijska²[\[https://orcid.org/0000-0002-1698-6826\]](https://orcid.org/0000-0002-1698-6826),
Susanne Arndt³[\[https://orcid.org/0000-0002-1019-9151\]](https://orcid.org/0000-0002-1019-9151), Sophia Leimer⁴[\[https://orcid.org/0000-0001-6272-204X\]](https://orcid.org/0000-0001-6272-204X),
Johanna Hickmann⁵[\[https://orcid.org/0000-0002-7535-8344\]](https://orcid.org/0000-0002-7535-8344), Marc Fuhrmans⁶[\[https://orcid.org/0000-0002-9826-018X\]](https://orcid.org/0000-0002-9826-018X)
and Giacomo Lanza⁵[\[https://orcid.org/0000-0002-2239-3955\]](https://orcid.org/0000-0002-2239-3955)

¹ Universität Stuttgart, Germany

² TU Braunschweig, Germany

³ Technische Informationsbibliothek, Germany

⁴ Universität Duisburg-Essen, Germany

⁵ Physikalisch-Technische Bundesanstalt, Germany

⁶ Technische Universität Darmstadt, Germany

Abstract. We present an approach to document research data in a human and machine readable way by creating JSON-LD metadata files based on the m4i ontology. m4i is based on top level ontologies and reuses concepts of widely accepted ontologies to embed information modelled in m4i in larger contexts like a knowledge graph connecting research data with projects, actors, methods, tools and publications. We use a real-life research example from the engineering domain to show how to describe a research process with its object of research, the different steps with input and output data, the actors, and the used methods and tools. The resulting metadata files can serve as low-threshold documentation in a file system, as an exchange format between tools, as an input for data repositories and as a source of information to be used by scripts and tools.

Keywords: Metadata, Ontology, Provenance tracking, Interoperability, JSON-LD, Documentation

1. Motivation

Metadata4Ing (m4i) is an ontology for a process-based description of research activities and their results, focusing on the provenance of both research data and material objects. In engineering and many other disciplines, there is no standard yet to describe data sets in a fully semantic way. Reasons are the lack of knowledge on how to employ ontologies in everyday research and the lack of tools supporting this process. But a lot of engineers – especially in scientific computing – are used to handle data in JSON formats. Particularly suitable for that purpose is JSON-LD [1], a format that adds semantic information to normal JSON data in form of a so called context file. Through the semantic context, metadata in JSON-LD files are machine readable and actionable thanks to the exact semantic identification of classes and properties. At the same time, the files remain readable and manageable for humans, at least for humans used to handle structured data in a JSON format. We will demonstrate this approach using a real-life research example.

2. Metadata4Ing at a Glance

m4i is intended as a general process model that allows a flexible description of research activities and their results applicable to many disciplines. As depicted in Figure 1, m4i offers a selection of unambiguous well-documented terms for general concepts like processing steps, in- and output, employed methods and tools that allows modelling information about research processes and results in a structured, consistent and machine-actionable way. All the terms are available on the NFDI4Ing Terminology Service [2].

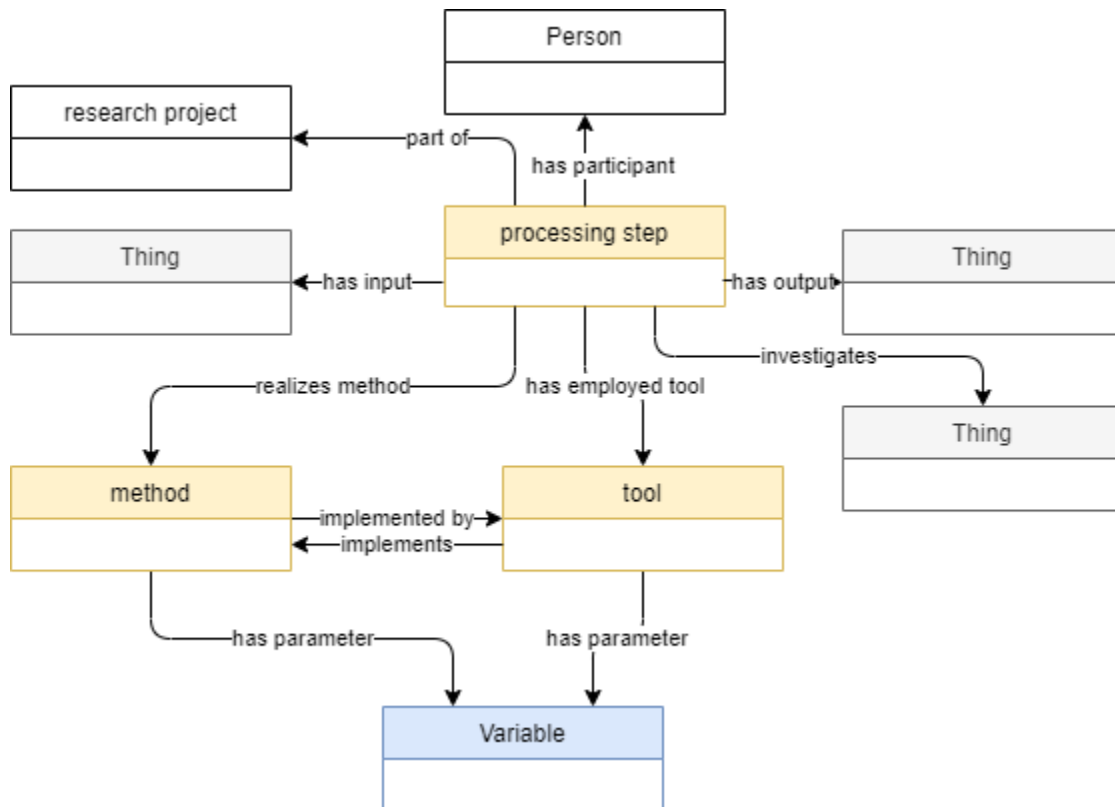


Figure 1: Core classes of m4i to describe scientific processes

One of the main benefits of using m4i is that the resulting description is highly interoperable and allows integration of data from very different scientific disciplines into a single knowledge graph. m4i has a high degree of compatibility with well-known top-level ontologies by deriving the main classes from the [Basic Formal Ontology \(BFO\)](#) [3], [schema.org](#) [4], the [PROV Ontology \(PROV-O\)](#) [5] and the [Data Catalog Vocabulary \(DCAT\)](#) [6], which seamlessly embeds information modelled in m4i in larger contexts. This approach is a prerequisite for [FAIR \(meta\)data](#) [7], especially for their [interoperability](#).

An overview of m4i classes and properties is available at the [ontology documentation](#) [8]. The ontology code is developed at [m4i's GitLab repository](#) [9] where also its [releases](#) are published, proposals for further development in the form of [issues](#) can be made, or information [how to contribute](#) to m4i can be found.

3. Modelling scientific processes with m4i

In the talk, we will use an example of a material examination by micro X-ray computed tomography, to show how a research process can be modelled with m4i resulting in a JSON-LD metadata file.

3.1. Example setting

In the example research process, a sample of a material (in this case asphalt concrete) is examined with an XRCT scanner. The whole process consists of four steps:

1. preparation and positioning of the sample, and configuration of the parameters,
2. data generation in form of the XRCT scan,
3. image processing with the help of reconstruction algorithms and
4. post processing of the data.

The experimental setup, consisting of a holder for precise positioning of the sample, an X-ray source and a detector, is described in detail by Ruf and Steeb [10]. A resulting dataset can be found in [11].

3.2. Easy application via human-readable metadata

We will show how to create a machine-readable JSON-LD file that documents the example research process with its object of research, the different steps with input and output data, the actors, and the used methods and tools. A central element is the use of a context file, which gives the metadata the semantic context that clearly defines what exactly is meant by the information in the file, but allows to refer to classes, properties or instances via their human-readable labels. That way, semantic modelling of scientific processes without deeper knowledge in ontologies and RDF becomes possible.

The resulting metadata files can serve, for example, as low-threshold documentation / description in a file system, as an exchange format between electronic lab books and data sets, as an input for data repositories and as a source of information to be used by scripts and tools.

This information is also available in a [first steps guide](#) [12] demonstrating this approach for people with an IT affinity, e.g. application developers, research software engineers, data stewards, and tech-savvy domain experts.

4. Conclusion

Advantages of documenting research data with m4i comprise that it contributes to the implementation of [good scientific practice](#) [13], makes use of consistent metadata when searching for, analysing or otherwise using the data, and benefits collaborative work. RDF metadata can be stored as [JSON-LD](#) [1], together with the research data or code it describes. This format offers semantically enriched information that is understandable by humans and machines. In addition, a machine-actionable documentation of the data also facilitates publishing or archiving data in data repositories in a citable way. Depending on the functionalities of a data repository, the metadata contained can be used to automatically create or update a dataset. Another benefit of semantic enrichment of data sets arises when included in a global knowledge graph. The structure of m4i allows to describe and connect datasets, persons, projects, methods and tools and therefore supports complex search queries, improving the retrieval of information, especially as data pools continue to grow and connect. With the first steps guide, we aim both at researchers who want to describe their research data in a sustainable way that makes use in a knowledge graph possible, but also at the developers of tools that embed the documentation of research processes and results in the scientific workflow.

Data availability statement

The ontology m4i is available on the NFDI4Ing Terminology Service <https://terminology.tib.eu/ts/ontologies/m4i> and is published together with documentation, examples and guides on:

S. Arndt, B. Farnbacher, M. Fuhrmans, S. Hachinger, J. Hickmann, N. Hoppe, M. T. Horsch, D. Iglezakis, A. Karmacharya, G. Lanza, S. Leimer, J. Munke, Johannes, D. Terzijska, J. Theissen-Lipp, C. Wiljes, & J. Windeck (2022). "Metadata4Ing: An ontology for describing the generation of research data within a scientific activity". (1.1.0). Zenodo. <https://doi.org/10.5281/zenodo.7706017>

Underlying and related material

The documentation of m4i is available under:

<https://w3id.org/nfdi4ing/metadata4ing/>

The context file of m4i is available under:

https://w3id.org/nfdi4ing/metadata4ing/m4i_context.jsonld

The first steps guide is available under:

<https://git.rwth-aachen.de/nfdi4ing/metadata4ing/metadata4ing/-/blob/1.1.0/training/first-steps-guide.md>

Author contributions

All authors contributed equally to the [conceptualization](#) of the work described and on the [review and editing](#) of the abstract. Dorothea Iglezakis [wrote the first draft](#) together with Džulija Terzijska, Sophia Leimer, Marc Fuhrmans and Johanna Hickmann.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors (Metadata4Ing Workgroup) would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.

Acknowledgement

The authors would like to thank Silvio Peroni for developing LODÉ, a Live OWL Documentation Environment, Daniel Garijo for developing Widoco and the Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine for developing Protégé.

References

1. M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, P.-A. Champin, N. Lindström (2020). "JSON-LD 1.1. A JSON-based Serialization for Linked Data". <https://www.w3.org/TR/json-ld11/> (accessed 2023-04-19).
2. "NFDI4Ing Terminology Service". <https://terminology.nfdi4ing.de/ts/> (accessed 2023-04-20)
3. B. Smith, P. Grenon, A. Ruttenberg, M. Brochhausen, W. Ceusters, M. Courtot, R. Di-pert, J. Hastings, C. Mungall, F. Neuhaus, D. Natale, N. Otte, J. Overton, B. Peters, R. Rudnicki, S. Schulz, S. Seppälä, H. Stenzhorn, J. Zheng (2020). "BFO Basic Formal Ontology". <https://basic-formal-ontology.org/> (accessed 2023-04-19).
4. "Schema.org". <https://schema.org/> (accessed 2023-04-19).
5. T. Lebo, S. Sahoo, D. McGuinness. "PROV-O: The PROV Ontology". W3C Recommendation 30 April 2013. <https://www.w3.org/TR/prov-o/> (accessed 2023-04-19).
6. D. Browning. "DCAT 2 Vocabulary". W3C. <https://www.w3.org/ns/dcat> (accessed 2023-04-19).
7. M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson; J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, vol. 3, pp: 160018. <https://doi.org/10.1038/sdata.2016.18>
8. Metadata4Ing Workgroup. "Metadata4Ing: An ontology for describing the generation of research data within a scientific activity". Revision 1.1.0. <https://metadata4ing.org/> (accessed 2023-04-19).
9. "metadata4ing". GitLab RWTH Aachen. <https://git.rwth-aachen.de/nfdi4ing/metadata4ing/metadata4ing> (accessed 2023-04-19).
10. M. Ruf, & H. Steeb. "An open, modular, and flexible micro X-ray computed tomography system for research". *Review of Scientific Instruments*, vol. 91 no.11, 2020, p. 113102. <https://doi.org/10.1063/5.0019541>
11. M. Ruf, & H. Steeb. "micro-XRCT data set of open-pored asphalt concrete." DaRUS. 2020. <https://doi.org/10.18419/darus-639>
12. "How to Use Metadata4Ing - First Steps Tutorial". <https://git.rwth-aachen.de/nfdi4ing/metadata4ing/metadata4ing/-/blob/1.1.0/training/first-steps-guide.md> (accessed 2023-04-20)
13. Deutsche Forschungsgemeinschaft. Guidelines for Safeguarding Good Research Practice. Code of Conduct. 2022. <https://doi.org/10.5281/zenodo.6472827>

Schema.org as a Lightweight Harmonization Approach for NFDI

Leyla Jael Castro¹[\[https://orcid.org/0000-0003-3986-0510\]](https://orcid.org/0000-0003-3986-0510), Juliane Fluck^{1,2}[\[https://orcid.org/0000-0003-1379-7023\]](https://orcid.org/0000-0003-1379-7023), Daniel Arend³[\[https://orcid.org/0000-0002-2455-5938\]](https://orcid.org/0000-0002-2455-5938), Matthias Lange³[\[0000-0002-4316-078X\]](https://orcid.org/0000-0002-4316-078X), Daniel Martini⁴[\[https://orcid.org/0000-0002-6953-4524\]](https://orcid.org/0000-0002-6953-4524), Steffen Neumann⁵[\[https://orcid.org/0000-0002-7899-7192\]](https://orcid.org/0000-0002-7899-7192), Sonja Schimmler⁶[\[https://orcid.org/0000-0002-8786-7250\]](https://orcid.org/0000-0002-8786-7250) and Dietrich Rebholz-Schuhmann^{1,7}[\[https://orcid.org/0000-0002-1018-0370\]](https://orcid.org/0000-0002-1018-0370)

¹ ZB MED Information Centre for Life Sciences, Cologne, Germany

² University of Bonn, Bonn, Germany

³ Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

⁴ Kuratorium für Technik und Bauwesen in der Landwirtschaft e.V. (KTBL)

⁵ Leibniz Institute of Plant Biochemistry, Halle, Germany

⁶ Fraunhofer Institute for Open Communication Systems (FOKUS), Berlin, Germany

⁷ University of Cologne, Cologne, Germany

Abstract. Schema.org is a controlled vocabulary that makes it easier for web pages to describe their actual content in a semantic, structured and machine-processable way. It is recognized by major search engines and data aggregators, making it easier for researchers to expose metadata describing their research outcomes. Here we present how Schema.org is used (or planned to be used) by some NFDI consortia, becoming a lightweight approach to harmonize digital objects coming from different sources so they can be connected to each other in a meaningful way.

Keywords: Metadata Harmonization, Lightweight Semantics, Schema.org, Metadata, Metadata Schema, Bioschemas

1. Background

Schema.org (from now on SchemaOrg) [1] is a vocabulary collaboratively developed by a community involving major search engines, including Google, Microsoft, Yahoo and Yandex. It offers a simple way for web pages to include structured data markup and thus semantically describe their content. Search engines can use that markup to present results tailored to the nature of the content, and offer added value to end-users. For instance, images are commonly displayed when looking for a recipe so users can get a graphic depiction; related recipes, e.g., including similar ingredients, can also be suggested. Structured markup also makes it possible to create summaries, like the ones displayed when looking for a movie which include similar movies, actors, release year, genre and more.

In recent years, the scientific community, with its ever increasing production of data, has shown interest in SchemaOrg as it presents low adoption barriers to publish data on the web [2]. While development of specialized APIs and web services requires software engineering skills, exposing SchemaOrg structured markup on web pages requires only basic understanding of HTML. As SchemaOrg is compatible with W3C RDF and Linked Data

specifications, the data described with it can be serialized using, for example, JSON-LD (current recommendation by SchemaOrg), and integrated into knowledge-graph-based infrastructures. Additionally, SchemaOrg types and properties can be reused within other RDF-based vocabularies. Another incentive for researchers to use SchemaOrg comes from the Google Data Search [3], a specialized portal released in 2020 helping researchers to find data on the web.

Selecting types and properties best suited to describe scientific outcomes is a different matter and will require some expertise on controlled vocabularies and semantics. Bioschemas [4] is a community project built on top of SchemaOrg, aiming to improve findability of resources in Life Sciences by embedding structured markup on relevant web pages. Bioschemas offers types tailored to Life Sciences but also profiles, i.e., usage recommendations including examples, on top of SchemaOrg types useful to describe scientific outcomes such as datasets, training materials, software and workflows. Other communities such as Science on Schema [5] and the Research Data Alliance Working Group Research Metadata Schemas [6], target particularly datasets and data catalogs.

2. Use of SchemaOrg in NFDI

Multiple NFDI consortia have turned to SchemaOrg as a lightweight approach to harmonize data from the different participant partners. As a side effect, they are also becoming connected along NFDI. Although at a basic level, SchemaOrg markup also contributes to make digital objects findable (via search engines or data aggregators and registries), accessible (exposed over TCP/IP protocol), interoperable (common types, properties and connections to each other), and reusable (via, e.g., inclusion of license and conditions of access).

Most consortia related to Life Science have turned to Bioschemas profiles as they already provide some guidance on how to use SchemaOrg in this domain. For instance, **FAIRagro** will build upon and extend Bioschemas specifications, taking also into account well-known vocabularies in the agri-domain (e.g., AgroVoc [7]). Work on extensions and adoption will involve a variety of domain experts, expert associations and service providers to work collaboratively, via two AgriHackathons. FAIRagro and DataPlant will also benefit from the work done by the ELIXIR Plant Community [8] which unites a diverse set of services and work on the different implementation studies to increase the Bioschemas compliance of their resources. There are already some first contacts initiated, which will be increased and intensified in the next few years. On its part, **NFDI4Microbiota** is starting with the integration of Bioschemas specifications related to training. **NFDI4Biodiversity** could also benefit from Bioschemas as it offers relevant types such as Taxon and TaxonName. Bioschemas also support chemical related types, MolecularEntity and ChemicalSubstance, that will be useful for **NFDI4Chem** and **NFDI4Cat**. The regular BioHackathons organized by e.g., ELIXIR or the German Node ELIXIR-DE provide opportunities to submit proposals and work on specific needs to improve metadata profiles, data resources or infrastructure.

Outside the Life Sciences domain, **NFDI4Culture** and **NFDI4MatWerk** have been collaborating to create a common ontology including elements from SchemaOrg [9]. Such an ontology can be easily extended to cover more domain-oriented terms, which has been done already in NFDI4MatWerk. Other consortia, such as **NFDI4DataScience** and **NFDI4Memory** are planning to pick up the approach. **NFDI4DataScience** is using SchemaOrg as the default representation for digital objects in their search engine and portal, including training datasets, artificial intelligence models (direct contribution as this object is not yet covered by SchemaOrg core), training and optimization software, and scholarly publications.

3. Future Work

SchemaOrg offers a broad spectrum of types and properties, some of them useful to represent research outcomes, some others that can be combined with domain-specific vocabularies and datasets. This ample coverage in SchemaOrg makes it difficult to use it in a consistent and coherent way (e.g., while someone can use free-text keywords, someone else could favor terms defined in a controlled vocabulary). Bioschemas profiles address this challenge by providing usage recommendations. Finding a way to harmonize across different NFDIs and avoiding duplication of efforts wrt new types and properties could become part of one of the projects in Base4NFDI. Broader adoption will require international acceptance. In the European context, ELIXIR Europe is one of the pioneers using SchemaOrg for science since 2017 [3, 10, 11, 12], while, at an international level, the Research Data Alliance has also contributed in this regard [6].

Data availability statement

No data was used to support the text presented in this abstract.

Author contributions

LJC: conceptualization, project administration, writing – original draft, writing – review & editing. JF, DA, ML, DM, STN, SS: writing - review & editing. DRS: conceptualization, funding acquisition, project administration, writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

NFDI consortia are funded by the Deutsche Forschungsgemeinschaft DFG: NFDI4DataScience project no. 460234259, FAIRAgro project no. 501899475, NFDI4Chem project no. 441958208.

References

1. Guha RV, Brickley D, Macbeth S (2016) Schema.org. *Communications of the ACM* 59 (2): 44- 51. <https://doi.org/10.1145/2844544>
2. Gray A, Castro LJ, Juty N, Goble C. (2023) Schema.org for Scientific Data. *Artificial Intelligence for Science*. World Scientific; pp. 495–514. https://doi.org/10.1142/9789811265679_0027
3. Benjelloun O, Chen S, Noy N. Google Dataset Search by the Numbers. 2020. <https://doi.org/10.48550/arXiv.2006.06894>
4. Gray AJG, Goble C, Jimenez RC (2017) From Potato Salad to Protein Annotation. ISWC Posters and Demo session. URL: <http://ceur-ws.org/Vol-1963/paper579.pdf>
5. Shepherd A et al. (2022). Science-on-Schema.org v1.3.0. Zenodo. <https://doi.org/10.5281/zenodo.6502539>
6. Wu M, Juty N, RDA Research Metadata Schemas WG, Collins J, Duerr R, Ridsdale C, et al. (2022). Guidelines for publishing structured metadata on the Web. RDA. <https://doi.org/10.15497/RDA00066>
7. Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., & Keizer, J. (2013). The AGROVOC Linked Dataset. *Semantic Web*, 4(3), 341–348. <https://doi.org/10.3233/SW-130106>

8. Pommier C, Gruden K, Junker A et al. (2021). ELIXIR Plant sciences 2020-2023 Roadmap. F1000Research 2021, 10(ELIXIR):145 <https://doi.org/10.7490/f1000research.1118482.1>
9. Tietz T, Bruns S, Sack H, Posthumus E. (version 1.1) NFDI4Culture Ontology. Available at <https://nfdi4culture.de/ontology>
10. García, L. J., Giraldo, O. L., Castro, A. G., & Dumontier, M. (2017). Bioschemas: schema. org for the Life Sciences. In SWAT4LS. <https://ceur-ws.org/Vol-2042/paper33.pdf>
11. Michel, F. (2018). Bioschemas & Schema. org: a lightweight semantic layer for life sciences websites. Biodiversity Information Science and Standards, 2, e25836. <https://doi.org/10.3897/biss.2.25836>
12. Castro LJ, Palagi PM, Beard N, Attwood TK, Brazas MD. (2022) Bioschemas Training Profiles: A set of specifications for standardizing training information to facilitate the discovery of training programs and resources. bioRxiv. <https://doi.org/10.1101/2022.11.24.516513>

Investigating the Landscape of Ontologies for Catalysis Research Data Management

Alexander S. Behr¹[\[https://orcid.org/0000-0003-4620-8248\]](https://orcid.org/0000-0003-4620-8248), Hendrik Borgelt¹[\[https://orcid.org/0000-0001-5886-7860\]](https://orcid.org/0000-0001-5886-7860),
Taras Petrenko²[\[https://orcid.org/0000-0002-0049-4835\]](https://orcid.org/0000-0002-0049-4835), Mark Dörr³[\[https://orcid.org/0000-0003-3270-6895\]](https://orcid.org/0000-0003-3270-6895),
Norbert Kockmann¹[\[https://orcid.org/0000-0002-8852-3812\]](https://orcid.org/0000-0002-8852-3812)

¹ Dept. of Biochemical and Chemical Engineering, TU Dortmund University, Germany

² High-Performance Computing Center Stuttgart (HLRS), University of Stuttgart, Germany

³ Dept. of Biotechnology & Enzyme Catalysis, University of Greifswald, Germany

Abstract:

This work provides a survey of ontologies for catalysis research to improve the findability, accessibility, interoperability, and reusability (FAIRness) of research data. Applying tools that are commonly used by lab scientists, ontologies relevant to catalysis research are classified in a simple, well formatted spreadsheet template (Excel). This enables a scientist and domain expert without programming skills to evaluate a certain ontology. The entries of this template are then processed and visualized through automated creation of markdown files on GitHub using Python scripts. Furthermore, ontology mapping by searching for similar pairs of classes across different ontologies is performed, using the outcome of the ontology classification. This work contributes to the development of ontologies for catalysis research, facilitating better data integration and knowledge sharing while reusing existing semantic artefacts.

Keywords: Ontology Collection, Catalysis, Semantic Web, Classification

1. Introduction

As digitization of the scientific community advances, the need for FAIR (Findable, Accessible, Interoperable, Reusable) data rises to ensure machine-processability of data. Enabling a higher data FAIRness, ontologies represent knowledge explicitly in a machine-understandable way. [1] Furthermore, research data occurring in the field of catalysis research often is complex and diverse. Thus, the NFDI4Cat consortium focuses on ontology development for the catalysis research domain. [2]

To enhance semantic interoperability and compliance with existing ontologies, a collection of ontologies and semantic artefacts was created with importance to the data value chain of catalysis research. [3] Some of these ontologies are not easily reusable and do not provide proper documentation. This work presents a reiteration of the initial ontology landscape for catalysis research. The workflow and software is developed to be as reusable as possible, to enable other domains for such ontology classification.

2. Methods

To identify suitable ontologies, ontologies listed in the OLS [4] and BioPortal [5] are screened by look-up of different keywords. Additionally, the ontologies listed in [3] and [6] are considered.

The ontology survey is conducted with the help of a well formatted and intuitively designed spreadsheet template (Excel) to simplify access and handling of the ontology collection, capturing the relevant information on each ontology. For each ontology, such a template is filled in consisting of five topics and a comment section listed in Table 1 along the exact content included in each topic.

Table 1. Classification scheme of the ontologies. Information regarding the five topics is gathered for each ontology to classify the ontologies regarding the content of each topic.

Topic	Content
General information on the ontology	Ontology name, alternative names, ontology acronym, creator(s) & issuing organization, kind of organizational structure
References	Organizational website, persistent URI of ontology file, link to documentation, link to version directory, additional links
Ontology modeling and availability	Provided ontology formats (ttl, owl, ...), degree of inference and composition (inferred, non-inferred, compacted, ...), license, working reasoners, shortest reasoning time, alignment with TLO, ontology imports, prefixes used, class annotation types
Classification of contained domains of interest	Biocatalysis, heterogenous catalysis, homogenous catalysis, chemical substance modeling, material modeling, process modeling, synthesis data, operando data, performance data, characterisation data, heat, transport and kinetic data, process design, energy and cost data, top level ontology
Ontology characteristics	Axioms, logical axiom count, declaration, class count, object property count, data property count, individual count, annotation property count
Comments	Any additional comments or remarks on topics not covered by the other topics

To facilitate documentation and access, the content of the Excel file is used to automatically generate markdown files, that contain simplified, text-based formatting instructions and can be rendered similarly to HTML. Rendering the markdown files in GitHub provides a comprehensive and interactive overview of each ontology, making it easier for researchers to assess the suitability of an ontology for their research needs. The data is then imported by Python to generate JSON files increasing machine-readability of the results, which eases further use of the data.

Overall, the Excel template and automated generation of markdown and JSON files enable efficient data collection, documentation, and access. This helps in automated detection of similar classes (mapping) between ontologies, too. For this, a Python script is used that detects similarities of ontology classes of two ontologies based on similar labels, prefLabels, altLabels, class names, and IRIs.

3. Results

The survey in this contribution classifies 28 ontologies containing the data as listed in Table 1. After conversion of the Excel-file to markdown files for the respective ontologies, the markdown files are visualized in the browser page of the GitHub repository. This also allows for a simple and clear presentation of the data, accessible without restrictions. The repository landing page contains a readme-file listing some general information about the ontology collection. Additionally, the acronym and the name of each ontology are listed in a table. As the markdown file allows for linking between different files, clicking on an acronym of an ontology directly redirects to the respective markdown file within the repository. The respective opened markdown file is visualized, too, and lists the information as given in Table 1. Figure 1 depicts the visualization

of the repository readme file (left) and a resulting ontology information page (right) of the ChEBI ontology.

Ontology World Map of NFDI4Cat

Repository which lists ontologies relevant for catalysis research.

For remarks, additions, or general questions either use the issues or contact the responsible person (see below). For contributions please download the markdown file called [General Template](#) and contact us either via mail, issue or pull request with your updated markdown file. A condensed view on the data provided in the markdown-files is given in [master_table](#). The respective markdown files for each ontology listed in the table below are located in [ontology_metadata](#). In the subdirectory [json](#), the information contained for each ontology is stored in json-format for an ease in access of the data presented in markdown.

Contact: alexander.behr@tu-dortmund.de

Ontology Metadata files

These are the ontologies and links to the ontology markdown files, NFDI4Cat deems as relevant:

Link to Markdown	Ontology Name
AFO	Allotrope Foundation Ontology
BFO	Basic Formal Ontology
CAO	Chemical Analysis Ontology
CHEBI	Chemical Entities of Biological Interest
CHEMINF	Chemical Information Ontology
CHMO	Chemical Methods Ontology
EMMO	Elementary Multiperspective Material Ontology
ENVO	The Environment Ontology
OntoCAPE	Ontology for the domain of Computer Aided Process Engineering
OSMO	Ontology for Simulation, Modelling, and Optimization
REX	Ontology on Physico-chemical Processes

ChEBI - Chemical Entities of Biological Interest

Ontology

Aspect	Description
Full Name	Chemical Entities of Biological Interest
Synonyms/Alternative Names	chebi_ontology
Ontology Acronym	ChEBI
Creator(s) & Issuing Organisation	Michael Ashburner & Pankaj Jaiswal.
Nature of Organisational Structure	ChEBI curation team

References

Aspect	Description
Organisational Website	https://www.ebi.ac.uk/chebi/int.do
Persistent URI of Ontology File	http://purl.obolibrary.org/obo/chebi.owl
Link to Documentation	Documentation available at organisational website (user manual, annotation manual, developer manual), but seem to have no permalinks, but are google documents
Link to Version directory	https://ftp.ebi.ac.uk/pub/databases/chebi/ontology/
Optional links (Papers, Repos...)	Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. <i>Nucleic Acids Res.</i> More info on available formats etc.: https://www.ebi.ac.uk/chebi/downloadsForward.do

Ontology Modeling And Availability

Aspect	Description
Ontology Formats Provided	sdf; owl; obo; flat file; Oracle binary table dump; SQL table dump
Degree of Inference/Composition	not defined
License	Creative Commons 4.0 (CC BY 4.0)
Validated Reasoning with	HermiT
Shortest reasoning time	128730 ms
Aligned with Top Level Ontology	OBO
Imports Ontology(ies)	only self
Prefixes used	obo:chebi:xsd; rdfs:xml; rdf:owl; obo:inOwl; chebi1; chebi2; chebi3; chebi4
Class annotation types	rdfs:Label; obo:Definition (IAO_0000115)

Figure 1. Visualization of the ontology classification via markdown files on GitHub. The repository readme file (left) lists the ontologies and links to the markdown files describing the respective ontology (right) according to the classification listed in Table 1.

After classification of the ontologies, the search for class similarities is performed automatically for each pair of ontologies. This helps to identify close ontologies and get common classes to extend existing ontologies. The resulting list of common classes for each pair of ontologies is depicted in Figure 2 as heat map with low count of classes in red and high count of classes in green for 11 ontologies. The CHEBI ontology has, for example, 937 common classes with the ENVO ontology, which is the intersection of the two largest ontologies.

	AFO	BFO	CAO	CHEBI	CHMO	EMMO	ENVO	OSMO	REX	SBO	VIMMP
AFO	3028										
BFO	35	36									
CAO	27	2	446								
CHEBI	43	0	39	182375							
CHMO	232	12	53	22	3102						
EMMO	2	0	5	2	0	199					
ENVO	158	26	59	937	32	2	6997				
OSMO	6	0	0	0	0	0	0	173			
REX	8	0	2	0	18	0	6	0	553		
SBO	26	2	3	12	3	0	14	1	11	695	
VIMMP	40	2	12	2	3	25	16	7	0	8	1082

Figure 2. Heatmap of the 11 ontologies investigated. Green entries show a high absolute number of common classes, while red indicates the contrary.

Data availability statement

The data, code and markdown files presented in this abstract will be available at GitHub here: <https://github.com/AleSteB/Ontology-Overview-of-NFDI4Cat>

Author contributions

Conceptualization: A.S.B., H.B., T.P., M.D.; Methodology: A.S.B., H.B.; Software: A.S.B.; Validation: A.S.B.; Data Curation: A.S.B., H.B.; Writing – Original Draft: A.S.B., Writing – Review & Editing: N.K., M.D., T.P.; Visualization: A.S.B.; Supervision: A.S.B., N.K.

Competing interests

The authors declare that they have no competing interests.

Funding

The Deutsche Forschungsgemeinschaft (DFG) is acknowledged for funding this research as part of the Nationale Forschungsdateninfrastruktur (NFDI) initiative (grant No.: NFDI/2-1 - 2021).

Acknowledgement

The authors would like to thank all members of the NFDI4Cat Task Area 1.

A.S.B. thanks the networking program ‘Sustainable Chemical Synthesis 2.0’ (SusChemSys 2.0) for the support and fruitful discussions across disciplines.

References

1. T. R. Gruber, “A translation approach to portable ontology specifications,” *Knowl.Acquis.*, vol. 5, no. 2, pp. 199–220, 1993, doi: <https://doi.org/10.1006/knac.1993.1008>
2. C. Wulf et al., “A Unified Research Data Infrastructure for Catalysis Research – Challenges and Concepts,” *ChemCatChem* 2021, 13, 3223 doi: <https://doi.org/10.1002/cctc.202001974>
3. M. Horsch et al., “Interoperability and Architecture Requirements Analysis and Metadata Standardization for a Research Data Infrastructure in Catalysis,” In: Pozanenko, A., Stupnikov, S., Thalheim, B., Mendez, E., Kiselyova, N. (eds) *Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2021. Communications in Computer and Information Science*, vol 1620. Springer, Cham., 2022, doi: https://doi.org/10.1007/978-3-031-12285-9_10
4. S. Jupp et al., “A new Ontology Lookup Service at EMBL-EBI,” In: Malone, J. et al. (eds.) *Proceedings of SWAT4LS International Conference 2015*, URL: <https://www.ebi.ac.uk/ols/index>
5. N.F. Noy et al., “BioPortal: ontologies and integrated data resources at the click of a mouse,” *Nucleic Acids Res.* 2009 Jul 1;37(Web Server issue):W170-3. Epub 2009, URL: <https://biportal.bioontology.org/>
6. P. Strömert et al., “Ontologies4Chem: the landscape of ontologies in chemistry,” *Pure and Applied Chemistry*, vol. 94, no. 6, 2022, pp. 605-622. <https://doi.org/10.1515/pac-2021-2007>

Provenance Core Data Set

A Minimal Information Model for Data Provenance in Biomedical Research

Ulrich Sax^{1,5}[\[https://orcid.org/0000-0002-8188-3495\]](https://orcid.org/0000-0002-8188-3495), Christian Henke¹[\[https://orcid.org/0000-0002-4541-4018\]](https://orcid.org/0000-0002-4541-4018), Christian Draeger²[\[https://orcid.org/0000-0001-8835-4548\]](https://orcid.org/0000-0001-8835-4548), Theresa Bender¹[\[https://orcid.org/0000-0001-6721-7034\]](https://orcid.org/0000-0001-6721-7034), Alessandra Kuntz^{1,5}[\[https://orcid.org/0000-0002-8259-2577\]](https://orcid.org/0000-0002-8259-2577), Martin Golebiewski^{3,5}[\[https://orcid.org/0000-0003-2039-8733\]](https://orcid.org/0000-0003-2039-8733), Hannes Ulrich⁴[\[https://orcid.org/0000-0002-8349-6798\]](https://orcid.org/0000-0002-8349-6798) and Matthias Löbe^{2,5}[\[https://orcid.org/0000-0002-2344-0426\]](https://orcid.org/0000-0002-2344-0426)

¹ Department of Medical Informatics, University Medical Center Göttingen, Germany

² University Leipzig, Germany

³ Heidelberg Institute of Theoretical Studies, Heidelberg, Germany

⁴ Institute for Medical Informatics and Statistics, Kiel University, Germany

⁵ part of the NFDI4Health Consortium

Abstract. The exchange, dissemination, and reuse of biological specimens and data have become essential for life sciences research. This requires standards that enable cross-organizational documentation, traceability, and tracking of data and its corresponding metadata. Thus, data provenance, or the lineage of data, is an important aspect of data management in any information system integrating data from different sources [1]. It provides crucial information about the origin, transformation, and accountability of data, which is essential for ensuring trustworthiness, transparency, and quality of healthcare data [2]. For biological material and derived data, a novel ISO standard was recently introduced that specifies a general concept for a provenance information model for biological material and data and requirements for provenance data interoperability and serialization [3,4]. However, a specific standard for health data provenance is currently missing. In recent years, there has been a growing need for developing a minimal core data set for representing provenance information in health information systems. This paper presents a Provenance Core Data Set (PCDS), a generalized data model that aims to provide a set of attributes for describing data provenance in health information systems and beyond.

Keywords: data provenance, lineage, Life Sciences, Harmonizing RDM, Linking RDM

1. Methods

The Provenance Core Data Set was developed based on inputs from various web conferences and discussions among experts in the field of health informatics organized by the NMDR2 project. Several data and metadata standards were examined on their ability to capture provenance metadata. The data model focuses on general attributes that are applicable to different scenarios and use cases, including data distribution, data transformation, and accountability.

2. Results

The Provenance Core Data Set provides a simple data model for representing data provenance in health information systems. It includes attributes that can capture important aspects

of provenance, such as the time of data creation, modification, and update, the source system information, the status of the data, accuracy assessment, and responsible parties. The data model is intended to support the trustworthiness, transparency, and accountability of data in health information systems, which are essential for ensuring data quality and integrity.

The data model includes attributes such as create date, change date, update date, source system type, source system name, source system URL, source system release, source system vendor name, source system vendor URL, status, accuracy, creator, interpretive comment, provider, frequency, depends on, measurement method, and measured by. These attributes are intended to provide comprehensive information about the provenance of data in health information systems.

3. Discussion

The Provenance Core Data Set can be applied to different scenarios and use cases in health information systems. It provides a common set of attributes that can be used to describe data provenance in a standardized and consistent manner. The data model can be used to capture important information about the origin, transformation, and accountability of data, which can be useful for various purposes such as data quality assessment, data integration, and data sharing. However, further research and validation are needed to evaluate the applicability and effectiveness of the Provenance Core Data Set in real-world health information systems.

4. Outlook

The Provenance Core Data Set is a promising approach for representing data provenance in health information systems and beyond. We need to discuss this approach with other communities especially in the NFDI context.

The data model has the potential to support the trustworthiness, transparency, and accountability of data in health information systems, which are crucial for ensuring data quality and integrity. Further research and validation are needed to evaluate the applicability and effectiveness of the Provenance Core Data Set in different health information systems and beyond. More standards integration has to be organized regarding huge initiatives like the German Medical Informatics Initiative [5] and their HL7 FHIR based core data set [6,7].

Data availability statement

-

Underlying and related material

-

Author contributions

All authors collaborate in the NFDI4health project. US prepared the manuscript, all authors reviewed and finalized.

Competing interests

The authors declare that they have no competing interests.

Funding

We greatly appreciate the funding from the Deutsche Forschungsgemeinschaft (DFG) through projects no. 442326535 (NFDI4health), 451265285 (NFDI4health TF COVID19), and 315072261 (NMDR2).

Acknowledgement

-

References

1. L. Moreau, and P. Missier, PROV-DM: The PROV Data Model, (2013). <https://www.w3.org/TR/prov-dm/> (accessed April 24, 2023)
2. Parciak et al 2019: Provenance Solutions for Medical Research in Heterogeneous IT-Infrastructure: An Implementation Roadmap; Studies in Health Technology and Informatics Volume 264: MEDINFO 2019: Health and Wellbeing e-Networks for All; DOI 10.3233/SHTI190231[3] R. Wittner, P. Holub, C. Mascia, F. Frexia, H. Müller, M. Plass, C. Allocca, F. Betsou, T. Burdett, I. Cancio, A. Chapman, M. Chapman, M. Courtot, V. Curcin, J. Eder, M. Elliot, K. Exter, C. Goble, M. Golebiewski, B. Kisler, A. Kremer, S. Leo, S. Lin-Gibson, A. Marsano, M. Mattavelli, J. Moore, H. Nakae, I. Perseil, A. Salman, J. Sluka, S. Soiland-Reyes, C. Strambio-De-Castillia, M. Sussman, J.R. Swedlow, K. Zatloukal, J. Geiger, "Toward a common standard for data and specimen provenance in life sciences", Learn Health Sys., e10365, April 2023, doi: <https://doi.org/10.1002/lrh2.1036>
4. "ISO/TS 23494-1:2023 Biotechnology — Provenance information model for biological material and data — Part 1: Design concepts and general requirements" <https://www.iso.org/standard/80715.html> (accessed 25 April 2023) [5] Semler et al 2018: German Medical Informatics Initiative; Methods Inf Med. 2018 Jul; 57(Suppl 1): e50–e56.PMID: 30016818
5. The Medical Informatics Initiative Core data set: <https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set> (accessed April 23, 2023)7 FHIR Provenance: <https://fhir-ru.github.io/provenance.html> (accessed April 23, 2023)

FAIR Data for Energy System Research

An Overview of NFDI4Energy Task Area 4

Amanda Wein¹[\[https://orcid.org/0009-0009-2960-3474\]](https://orcid.org/0009-0009-2960-3474), Jan Reinkensmeier¹, Anke Weidlich²[\[https://orcid.org/0000-0003-2361-0912\]](https://orcid.org/0000-0003-2361-0912), Johan Lilliestam³[\[https://orcid.org/0000-0001-6913-5956\]](https://orcid.org/0000-0001-6913-5956), Veit Hagenmeyer⁴[\[https://orcid.org/0000-0002-3572-9083\]](https://orcid.org/0000-0002-3572-9083), Mascha Richter⁵, Sören Auer⁶[\[https://orcid.org/0000-0002-0698-2864\]](https://orcid.org/0000-0002-0698-2864), Astrid Nieße⁷[\[https://orcid.org/0000-0003-1881-9172\]](https://orcid.org/0000-0003-1881-9172), and Sebastian Lehnhoff¹[\[https://orcid.org/0000-0003-2340-6807\]](https://orcid.org/0000-0003-2340-6807)

¹ OFFIS e.V., Germany

² Albert-Ludwigs-Universität Freiburg, Germany

³ Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

⁴ Karlsruhe Institute of Technology – Institute for Automation and Applied Informatics, Germany

⁵ Reiner Lemoine Institut gGmbH, Germany

⁶ Technische Informationsbibliothek, Germany

⁷ Carl von Ossietzky Universität Oldenburg, Germany

Abstract. The NFDI4Energy consortium aims to establish new services filling a variety of needs for the energy system research community, from making FAIR research data easily accessible to promoting collaboration among community entities. Seven Task Areas (TAs) have been defined to achieve the consortium’s objectives, each with a specific focus. Task Area 4 (TA4): FAIR Data for Energy System Research shall develop ontologies, metadata standards, and services to promote semantic consistency and improve interoperability of energy research projects, thereby supporting the harmonization of data management among various institutions and research fields.

Keywords: Energy System Research, FAIR Data, Ontologies, Metadata, Data Infrastructure

1. Motivation

Energy system research is a highly interdisciplinary field, relying heavily upon expertise and existing models from engineering, economics, and meteorology, among other areas. The sheer quantity of data required for energy system models, and the disparate sources of this data, challenge the efficiency of energy research projects. Standardization of research data and metadata according to the FAIR Principles [1] is expected to yield significant benefits, including improved cooperation among research entities and greater ease in accessing relevant data sets for projects [2].

To promote and enable the usage of FAIR data in energy system research, TA4 focuses on several objectives:

- creation of ontologies for energy system research,
- creation of metadata standards to cover relevant Digital Objects (DOs), and
- implementation of a metadata registry and persistent identifier (PID) service for DOs [2].

The services developed under TA4 will form a semantic layer for the overall NFDI4Energy platform. This TA is a central component of the consortium's goals, supported by and providing support to multiple other TAs. Input from TA1 (Building and Serving the Energy Research Community) and TA2 (Integrating Society and Policy in Energy Research) will be key to ensuring the relevance of the ontology and metadata standards produced by TA4. TA5 (Simulation in Interdisciplinary Energy Research) will develop an energy simulation software ontology and a simulation model registry, which must be capable of full integration with the domain ontology and metadata registry developed by TA4. In addition, TA6 (Use Cases for Community Services) will define use cases for the services developed by NFDI4Energy, based on TA4's methods and services. [2]

2. Methodology

2.1 Ontologies & Standards

Two related ontologies shall be worked on. The first, a domain ontology, shall establish a common vocabulary for energy system research and shall hierarchically store knowledge from this domain; it will indicate relationships between concepts and will be capable of linking to the ontologies of other research areas, enabling interdisciplinary applications. The second ontology will extend the domain ontology by establishing terminology for defining long-term energy system scenarios. This scenario ontology shall allow for better comparisons between scenarios, based on the vocabulary established by the domain ontology. [2]

Several well-known ontologies for the energy domain currently exist, such as the Open Energy Ontology [3]. Therefore, the approach taken by TA4 will start with assessing these ontologies to determine which one(s) may be optimal for extending into the new domain and scenario ontologies. This ontology selection, improvement, and extension will be based upon an examination of features needed in the new ontologies and a comparison to features available in existing ontologies. Continuous feedback from other TAs and from researchers throughout the development process will ensure that the new ontologies closely align with the needs of the scientific community.

Together with the ontologies, a set of metadata standards will be created using the domain ontology. TA1, TA2, and TA6 will assist with defining requirements for these standards; in addition, TA4 will collaborate with the other NFDI consortia to discuss best practices for designing metadata standards that encourage interoperability with other standards. [2]

Working groups shall be established to encourage the scientific community's participation in this ontology and standards development effort [2].

2.2 Infrastructure & Services

Several software services are planned to support TA4's ontologies and standards, as well as the FAIRification of DOs.

A Terminology Service (TS) based on W3C standards will function as the main access point to search for and retrieve terminology information. It will include a registry tool to enable the registering in the TS of not only the TA4 domain and scenario ontologies, but also additional ontologies which are relevant for energy systems research. [2]

An Open Research Knowledge Graph (ORKG) [4] for energy research scenarios is also planned, as a service to simplify scenario comparison. The ORKG information structure is designed to semantically link knowledge from various research publications, making this knowledge both human- and machine-readable. In this context, the graph shall contain data on scenarios defined according to the scenario ontology. [2]

A metadata registry shall be developed based on the Leibniz Data Manager (LDM) [5]. This tool will link digital repositories of research data, allowing for visualizations and metadata maintenance of data sets. It shall assist researchers in selecting appropriate data sets for their projects, and improve the interoperability of data sets through standardized metadata. [2]

PIDs are a key element of FAIR data [6], and therefore a key element of TA4's work. A PID service will be created to interface with the metadata registry, using either an existing PID concept or, if necessary to meet user needs, a new domain-specific concept. All PIDs will be connected in a PID graph [7] to improve DO findability. [2]

Finally, several services shall facilitate the integration of TA4's artifacts with the overall NFDI4Energy platform: federated access and search services, REST APIs, and a submission service for publishing DOs. [2]

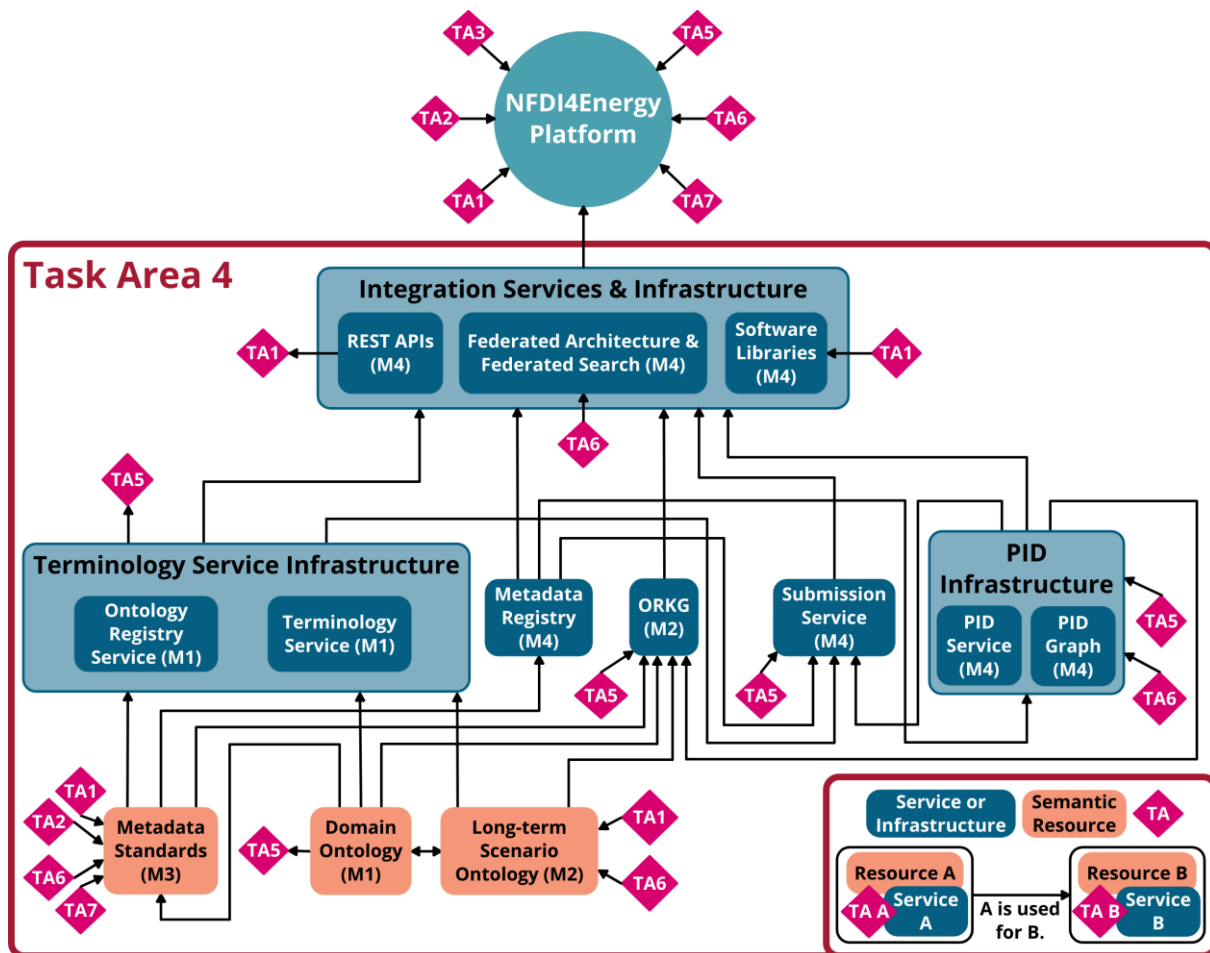


Figure 1. An overview of the resources to be created by TA4, with arrows indicating links between resources. The goals of TA4 are internally called Measures and are noted here as M1-M4.

3. Participants

TA4 will be led by OFFIS, supported by TA Lead Sebastian Lehnhoff and TA Coordinator Amanda Wein. Organizations contributing to TA4 are Albert-Ludwigs-Universität Freiburg, Friedrich-Alexander-Universität Erlangen-Nürnberg, Karlsruher Institute of Technology – Institute for Automation and Applied Informatics, Reiner Lemoine Institut, Technische Informationsbibliothek, and Carl von Ossietzky Universität Oldenburg. [2]

Competing interests

The authors declare that they have no competing interests.

Funding

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project 501865131.

References

1. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, article 160018, Mar. 2016, doi: <https://doi.org/10.1038/sdata.2016.18>.
2. A. Nieße, S. Ferenz, S. Auer, S. Dähling, S. Decker, J. Dorfner, et al. "nfdi4energy – National Research Data Infrastructure for the Interdisciplinary Energy System Research." Zenodo. <https://zenodo.org/record/6772013#.ZEYuontBybg> (accessed Apr. 24, 2023). doi: <https://doi.org/10.5281/zenodo.6772013>.
3. M. Booshehri, L. Emele, S. Flügel, H. Förster, J. Frey, U. Frey, et al., "Introducing the Open Energy Ontology: Enhancing data interpretation and interfacing in energy systems analysis," *Energy and AI*, vol. 5, article 100074, Sept. 2021, doi: <https://doi.org/10.1016/j.egyai.2021.100074>.
4. S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. E. Farfar, et al., "Improving Access to Scientific Literature with Knowledge Graphs," *Bibliothek Forschung und Praxis*, vol. 44, no. 3, pp. 516-529, 2020, doi: <https://doi.org/10.1515/bfp-2020-2042>.
5. "Leibniz Data Manager." TIB Labs. <https://labs.tib.eu/info/en/project/leibniz-data-manager/> (accessed Apr. 18, 2023).
6. "Persistent Identifiers (PIDs)." PID Kompetenzzentrum: Ein Service der TIB. <https://projects.tib.eu/pid-service/en/persistent-identifiers/persistent-identifiers-pids/> (accessed Apr. 18, 2023).
7. M. Fenner and A. Aryani. "Introducing the PID Graph." DataCite Blog. <https://blog.datacite.org/introducing-the-pid-graph/> (accessed Apr. 18, 2023). doi: <https://doi.org/10.5438/jwvf-8a66>.

Towards FAIR Research Data in Metrology

Giacomo Lanza ¹ [<https://orcid.org/0000-0002-2239-3955>], Martin Koval ² [<https://orcid.org/0000-0002-6360-1060>], Jean-Laurent Hippolyte ³ [<https://orcid.org/0000-0002-5263-2881>], Maitane Iturrate-Garcia ⁴ [<https://orcid.org/0000-0002-1517-3305>], Olivier Pellegrino ⁵ [<https://orcid.org/0000-0003-4167-1647>], Anne-Sophie Piette ⁶ [<https://orcid.org/0000-0002-7876-1251>], Federico Grasso Toro ⁷ [<https://orcid.org/0000-0002-9041-0868>]

¹ Physikalisch-Technische Bundesanstalt (PTB), Germany

² Český Metrologický Institut (CMI), Czech Republic

³ National Physical Laboratory (NPL), United Kingdom

⁴ Eidgenössisches Institut für Metrologie (METAS), Switzerland

⁵ Instituto Português da Qualidade (IPQ), Portugal

⁶ FPS Economy, National Standards (SMD), Belgium

⁷ GRASSO TORO Digital Solutions, Switzerland

Abstract. Good data management is necessary to maintain the trustworthiness and reliability of data. This is particularly important in metrology, the science of measurement, which ensures stable, comparable, coherent, and traceable measurement results. The digitalization of metrology has increased the demand for structured and harmonised research data management (RDM).

To meet this demand, the project TC-IM 1449 "Research data management in European metrology" was established in 2018. The project aims to promote good RDM practices underpinned by the FAIR principles, supporting traceability and reproducibility of measurement results. For that purpose, the project is providing researchers with the knowledge, competency, awareness, and tools to implement good RDM practices.

The project has formulated a vision for RDM in metrology for the support of scientists by developing and disseminating recommendations and in the organisation of training. As part of this vision, the project has produced several deliverables, including a template research data management policy, guidelines for data documentation, creation of metadata, and quality assurance for data publication. The project is also creating a comprehensive guide to RDM, a checklist for project coordinators, and providing training modules.

The project's activities reflect the needs of metrologists that are collated and communicated by the technical experts from the relevant Technical Committees and European Metrology Networks. Furthermore, the project's deliverables will be an invaluable resource for researchers seeking to effectively manage and share their research data.

Keywords: FAIR principles, Research data management, Metrology, Data management plans, Semantic technologies, Ontologies, Measurement units

1. Motivation

Metrology is the science of measurement, embracing both experimental and theoretical determinations of measurement uncertainty in science and technology [1]. Metrology

establishes a common understanding of measurement quantities and units, ensuring stable, comparable, coherent, and traceable measurement results. Moreover, it provides confidence in measurement at a stated level (usually described by a measurement uncertainty) and a structure based on scientific and technological concepts, thus underpinning all activities relying on accurate measurements.

For that purpose, it is essential to properly manage all data to make them trustable, reliable and protected against loss or corruption. Research data management (RDM) contributes to the digitalisation of metrological services, in line with both open science principles and the metrological principles of traceability and reproducibility. Indeed, both metrology and data management are prerequisites for efficient scientific research, which ensure that accurate and reliable information can be obtained from the produced data, leading to better understanding and reusability throughout science.

2. Project aims and background

To support this transition, the TC-IM 1449 project "Research data management in European metrology" was established in 2018. The project goal is fostering the development of harmonised practices for RDM of metrological data and services to improve the management of research data in metrology.

To this end, the project has formulated a vision for RDM in metrology and taken over the tasks to provide scientists with awareness, expertise and tools to implement it. This is achieved through the development and dissemination of recommendations in the form of guidelines, checklists, templates, and training.

The project's activities reflect the needs of metrologists that are collated and communicated by the technical experts from the European metrology consortium EURAMET. The engagement with scientific communities and other stakeholders is achieved through the EURAMET [Working Group on Metrology for Digital Transformation](#) (WG M4D) and ensures that guidance, training, and templates are relevant to metrologists, meet data management requirements by the relevant funders and communities, and create lasting data assets for the research community. All deliverables comply with the current "European Partnership on Metrology (EPM)" funding programme, which is based on Horizon Europe.

3. Project outputs

The project deliverables are designed to provide researchers with the tools and guidance they need to ensure their data are properly managed, stored, and shared. The co-dependencies between the project workstreams, inputs and outputs are illustrated in Figure 1.

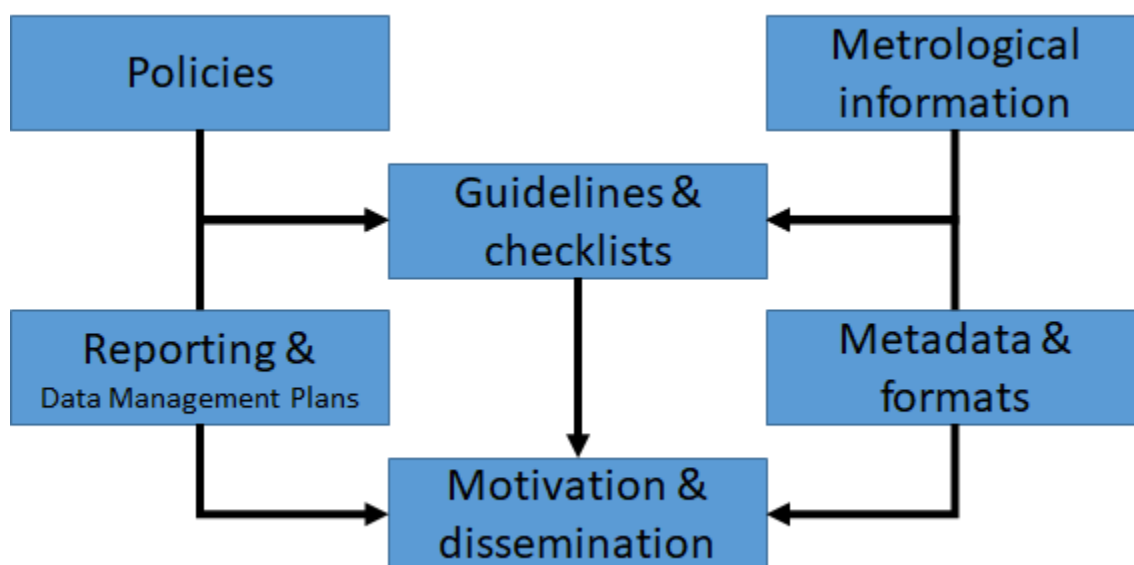


Figure 1. TC 1449 project workstreams, outputs and inputs.

In the first instance, we produced a template research data management policy. A wide adoption by metrology institutes, while catering to national specificities, would help harmonising the RDM practice and also offer a good example to other research institutions.

Moreover, we have produced a number of recommendations aiming to support researchers in doing good research data management during all different phases of a research activity. For the [funding application phase](#), we provide text snippets to fill in the sections of EPM project proposals about “Gender dimension”, “Open science” and “Research data management and management of other research outputs”. For the [project start](#), we provide a commented data management plan template, including controlled answer options and some help text for the most difficult parts.

To help researchers sharing and reusing data in machine-interpretable formats, guidelines for data documentation, creation of metadata and quality assurance for data publication are being prepared. Initial guidance on the adequate semantic representation of metrological information in a machine-actionable form has been published in an [article \[2\]](#) illustrating the benefits and limitations of open semantic technologies for metrology based on two real-world case studies. The key finding was the need for a flexible ontology of metrological concepts, agnostic to ontologies for quantities and units of measurement, thus facilitating community-driven good practices for FAIR and metrologically accurate representation of scientific data.

Further outputs include a comprehensive guide to RDM that will provide researchers with an overview of best practices and guidelines for managing their data throughout the research lifecycle, as well as a checklist for project coordinators to help align the project work with the best practices in RDM. The documents will be available by the end of 2023, providing a go-to resource for researchers who seek to improve their RDM.

To increase competences and awareness among our target group, a set of [training modules on research data management](#) for project coordinators was organised in 2021. Further training is planned in 2023 to support the preparation for the 2024 EPM funding call.

4. Partial conclusions and direct impact

This work presents the TC-IM 1449 project that aims to design and disseminate good RDM practices based on the FAIR principles as well as the core metrology concepts of traceability and reproducibility.

The project activities are integrated into the European metrology community framework for cooperation and knowledge sharing between all involved stakeholders. The project outputs relate to various stages of RDM and include definition of organisational policies, specification of metadata, elaboration of guidelines on data sharing and training of researchers. These outputs are aimed to support good RDM and accompany researchers and project coordinators in all phases of a research activity. The outputs are disseminated within the metrology community and beyond.

It is expected that TC-IM 1449 activities will lead to higher quality of research results and their documentation; reduction of costs for RDM; harmonised procedures for documentation of research projects; an easier dissemination of published data via repositories or databases; machine-interoperability of metrological information; improved compliance with the FAIR principles and funder's requests; and, as a result, higher chances of funding success.

5. Outlook

Future activities contemplate the definition of machine-interoperable editions of relevant terminology, such as the [International Vocabulary on Metrology \(VIM\)](#); recommendations for machine-interoperable data and metadata formats for metrological information (starting from the use case of key comparison data); recommendations for the design of repositories; checklists and guidelines for FAIRness maturity and FAIRification of datasets; recommendations for data quality evaluation and assurance (e.g., ISO 8000) [5]; procedures for semantic data validation in machine-to-machine data transfer. It is possible that the project outputs will contribute to the establishment of international standards (ISO, IEC, IEEE).

Data availability statement

Being a project about networking and capacity building, this project does not produce its own research data.

Underlying and related material

All project outputs (recommendations, training materials) are available online on the project's GitLab platform <https://gitlab1.ptb.de/GLanza/tc-im-1449> .

Author contributions

All authors contributed equally to the [conceptualization](#) of the work described and on the [review and editing](#) of the abstract. FGT [wrote the first draft](#) together with ASP, MIG, JLH, MK, OP, GL.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors are all EURAMET members of the (non-funded) project TC-IM 1449 "Research data management in European metrology".

Acknowledgement

The authors would like to thank the project's initiator, Sascha Eichstädt, all project partners contributing actively to our activities, all experts and stakeholders who provided us with valuable input, the past and present covenant of the M4D Workgroup, Sascha Eichstädt and Louise Wright, and the past and present Chairman of EURAMET's Technical Committee "Interdisciplinary Metrology", Miruna Dobre and Robert Gunn. We also would like to thank our project partner Alen Bošnjaković for useful corrections and review of the draft.

References

1. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. International vocabulary of metrology | Basic and general concepts and associated terms (VIM). Joint Committee for Guides in Metrology, JCGM 200:2012. (3rd edition). https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf/f0e1ad45-d337-bbeb-53a6-15fe649d0ff1
2. Hippolyte J-L, Romanchikova M, Bevilacqua M, Duncan P, Hunt SE, Grasso Toro F, Piette A-S, Neumann J. Using Ontologies to Create Machine-Actionable Datasets: Two Case Studies. *Metrology*. 2023; 3(1):65-80. <https://doi.org/10.3390/metrology3010003>
3. (GO FAIR) GO FAIR Initiative. FAIRification Process. Available online: <https://www.go-fair.org/fair-principles/fairification-process/>
4. VIM online: VIM definitions with informative annotations. Last update: 29 April 2017 <https://jcg.m.bipm.org/vim/en/index.html>
5. ISO 8000-1:2022 Data quality: <https://www.iso.org/standard/81745.html>

FAIR research data with NOMAD

FAIRmat's distributed, schema-based research-data infrastructure to harmonize RDM in materials science

Markus Scheidgen¹[\[https://orcid.org/0000-0002-8038-2277\]](https://orcid.org/0000-0002-8038-2277), Sebastian Brückner^{1,6}[\[https://orcid.org/0000-0002-5969-847X\]](https://orcid.org/0000-0002-5969-847X), Sandor Brockhauser¹[\[https://orcid.org/0000-0002-9700-4803\]](https://orcid.org/0000-0002-9700-4803), Luca M. Ghiringhelli¹[\[https://orcid.org/0000-0001-5099-3029\]](https://orcid.org/0000-0001-5099-3029), Felix Dietrich²[\[https://orcid.org/0000-0002-2906-1769\]](https://orcid.org/0000-0002-2906-1769), Ahmed E. Mansour¹[\[https://orcid.org/0000-0002-3411-6808\]](https://orcid.org/0000-0002-3411-6808), José A. Márquez¹[\[https://orcid.org/0000-0002-0640-0422\]](https://orcid.org/0000-0002-0640-0422), Martin Albrecht⁶[\[https://orcid.org/0000-0003-1835-052X\]](https://orcid.org/0000-0003-1835-052X), Heiko B. Weber⁷[\[https://orcid.org/0000-0002-6403-9022\]](https://orcid.org/0000-0002-6403-9022), Silvana Botti^{3,4}[\[https://orcid.org/0000-0002-4920-2370\]](https://orcid.org/0000-0002-4920-2370), Martin Aeschlimann⁵[\[https://orcid.org/0000-0003-3413-5029\]](https://orcid.org/0000-0003-3413-5029), and Claudia Draxl¹[\[https://orcid.org/0000-0003-3523-6657\]](https://orcid.org/0000-0003-3523-6657)

¹Physics Department and IRIS Adlershof, Humboldt-Universität zu Berlin, Germany

²School of CIT, Technical University of Munich, Munich, Germany

³Research Center Future Energy Materials and Systems of the University Alliance Ruhr, Faculty of Physics and Astronomy, Ruhr University Bochum, Bochum, Germany

⁴Institute for Solid State Theory and Optics, Friedrich Schiller University Jena, Jena, Germany.

⁵Department of Physics and Research Center OPTIMAS, University of Kaiserslautern, Kaiserslautern, Germany

⁶Leibniz-Institut für Kristallzüchtung, Berlin, Germany

⁷Department of Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Abstract: Scientific research is becoming increasingly data centric, which requires more effort to manage, share, and publish data. NOMAD is a web-based platform that provides research data management (RDM) for materials-science data. In addition to core RDM functionalities like uploading and sharing files, NOMAD automatically extracts structured data from supported file formats, normalizes, and converts data from these formats. NOMAD provides an extendable framework for managing not just files, but structured machine-actionable harmonized and inter-operable data. This is the basis for a faceted search with domain-specific filters, a comprehensive API, structured data entry via customizable ELNs, integrated data-analysis and machine-learning tools. NOMAD is run as a free public service and can additionally be operated by research institutes. Connecting NOMAD installations through the public services will allow a federated data infrastructure to share data between research institutes and further harmonize RDM within a large research domain such as materials science.

Keywords: materials science, research data management, electronic lab notebook, FAIR data, metadata

1 Introduction

In large research communities like materials science, researchers use many methods, instruments, tools, and workflows to produce large volumes of heterogeneous data files. The contained data describe semantically related research objects (like samples, materials, measurements) and it is believed that all combined data hold great potential for data re-use and artificial-intelligence-based solutions (therein including data mining and machine learning) [1], [2]. This is clearly being acknowledged not only by the research community but also by funding agencies, which are increasingly demanding coordinated efforts in research-data management (RDM) and the availability and longevity of open data by preserving and documenting all produced research data and metadata. This is the core mission of the German National Research Data Infrastructure (NFDI).

While individual researchers struggle with organizing and analyzing an increasing amount of data, communities face new challenges in making data Findable, Accessible, Interoperable, and Reproducible (FAIR) [3]. Collecting large amounts of heterogeneous files in generic repositories is not enough. Three key factors to FAIR data are (1) combining data with metadata and (2) putting all data into machine (and human) comprehensible and interoperable representations [4], [5], and (3) making the respective data-analysis tools accessible and let users directly execute them onto the stored data.

In this contribution, we present NOMAD, a distributed web-based platform for managing, sharing, and publishing FAIR data, based on rich domain-specific metadata and a homogeneous machine-comprehensible representation of data. NOMAD addresses RDM in two ways. First, it improves the data-centric workflows of individuals, laboratories, and research institutes by formalizing data acquisition, organizing and sharing data, homogenizing data for analysis, and integrating with analysis tools. This way, NOMAD provides the incentives and tools for research individuals to efficiently prepare FAIR (meta)data. Second, NOMAD allows researchers to collaborate on and publish harmonized data and serves communities as a repository for FAIR data.

2 NOMAD software and service

NOMAD allows users to upload and manage files similar to other data repositories, e.g., Zenodo [6]. Beyond that, NOMAD also processes uploaded files and extracts machine-actionable data from over 60 formats and atomistic-computation codes and over 70 experimental techniques and applications by supporting the NeXus standard [7]. All processed data follow a schema that defines how the data are hierarchically structured, cross-referenced, and aggregated. This schema is called the *NOMAD Metainfo*. It defines a general domain-independent superstructure, as well as highly detailed, specialized data from specific methods, tools, and programs. The NOMAD schema and processing can be extended by the community to support an ever-increasing number of file types. NOMAD simply provides the framework and interfaces to harmonize data from heterogeneous sources. Furthermore, the NOMAD schema can be used to define workflows and Electronic Lab Notebooks (ELN) to augment data from files with structured data entry by human researchers; see also Figure 1.

NOMAD can run containerized tools such as Jupyter notebooks directly on data. This allows users to implement, run, and share their analysis tools alongside the analyzed data. The ability to run analysis tools in the browser, without installation, and specifically configured for a respective dataset lets researchers not just share data and

analysis results, but the analysis itself and thereby drastically increases the reusability of data.

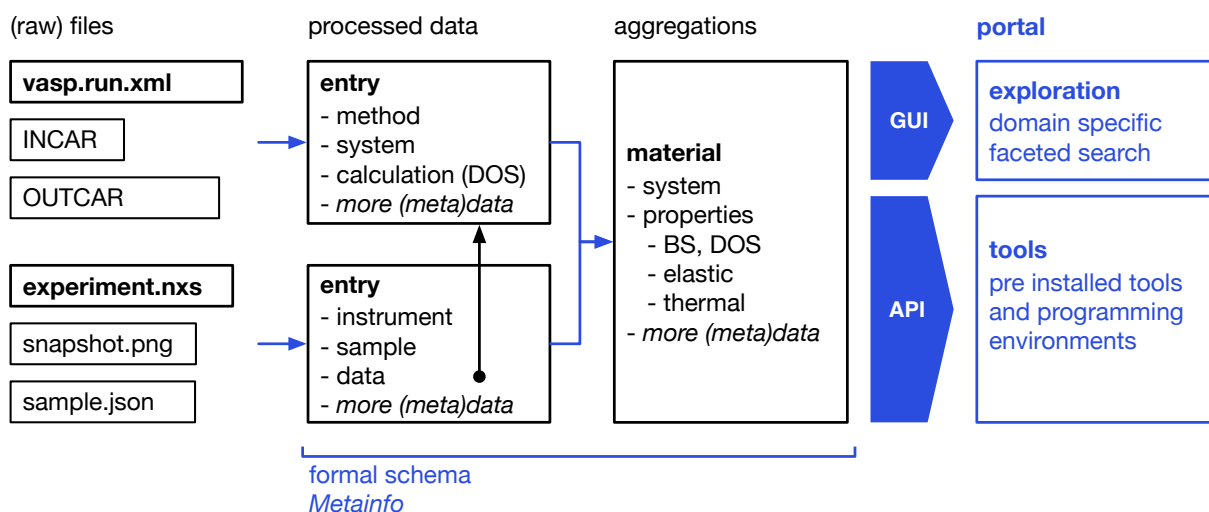


Figure 1. NOMAD provides a framework and interfaces for managing files alongside automatically extracted (meta)data.

NOMAD is built on open standards. For example, users can curate and publish datasets with a DOI and metadata based on DataCite [8], dataset metadata can be exported in various semantic web formats and is based on the DCAT vocabulary 2.0 [9]. NOMAD offers all computational results via the [OPTIMADE](#) [10] API specification. Other APIs are based on OpenAPI Specification [11]. Normalization of processed data is performed with widely accepted community software packages like MatID [12], ASE [13], or pymatgen [14].

The NOMAD software is used to operate a public and free [NOMAD service](#) that allows everyone to share and publish materials-science research data under the Create Commons Attribution License (cc-by) 4. This public NOMAD service contains as of April 2023 the data of over 12 million individual materials-science simulations and an increasing number of entries describing materials-science experiments. NOMAD is publicly available since 2014 and includes data from over 500 international authors. NOMAD includes the data of existing materials-science databases such as the [Materials Project](#) [15], [AFLOW](#) [16], [OQMD](#) [17], [EELSDB](#) [18], and the [Perovskite Database Project](#) [19], thereby overcoming heterogeneous database interfaces and providing all data with a singular API and in a common data representation.

The NOMAD software can also be independently operated by universities and other institutions to support local data management with independent data policies. Such self-managed installations are called *NOMAD Oases*, to distinguish them from the public NOMAD service. A NOMAD Oasis might be required when an institution needs to significantly customize the software for their needs, data volumes are too large to be conveniently transferred over the public internet, or if there are concerns about privacy or security. There will be the possibility to transfer data between different installations, and in order to adhere to the FAIR principles, the data (or at least metadata) in these Oases would ideally be made available to the public NOMAD service. NOMAD Oasis is used already by 15 research institutes and can be freely used under the Apache 2 open-source license.

The [NFDI consortium FAIRmat](#) develops NOMAD to build a federated FAIR-data infrastructure [1] (Figure 2). The NOMAD service will act as a central FAIRmat portal

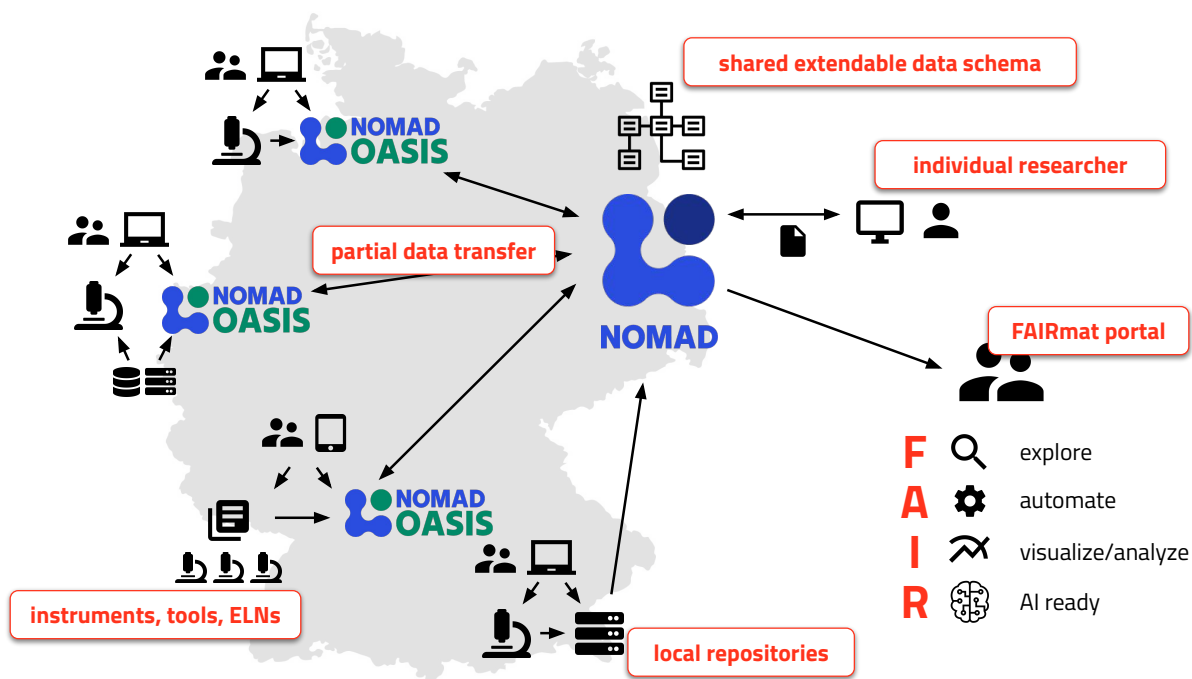


Figure 2. Distributed use of NOMAD in FAIRmat's federated data infrastructure.

that will allow for accessing the data managed in NOMAD installations of participating institutes. NOMAD's processing and shared data schema provide harmonization and interoperability, while its distributed nature ensures findability and accessibility across institutions. The potential to reuse and recontextualize data across a large heterogeneous community should provide the incentives to further extend the schema, processing, and data analysis to foster continuous data harmonization.

Funding

NOMAD software development is funded by the NFDI consortia FAIRmat (Deutsche Forschungsgemeinschaft, DFG, 460197019) and the NOMAD CoE (EU Horizon 2020, 951786); previous financial support was provided by the NOMAD CoE (EU Horizon 2020, 676580) and the Max-Planck Netzwerk BigMax. The Max Planck Computing and Data Facility (MPCDF) is hosting NOMAD's github and operating the public NOMAD service.

References

- [1] M. Scheffler, M. Aeschlimann, M. Albrecht, *et al.*, "Fair data enabling new horizons for materials research," *Nature*, vol. 604, no. 7907, pp. 635–642, 2022. DOI: [10.1038/s41586-022-04501-x](https://doi.org/10.1038/s41586-022-04501-x).
- [2] L. Sbailò, Á. Fekete, L. M. Ghiringhelli, and M. Scheffler, "The nomad artificial-intelligence toolkit: Turning materials-science data into knowledge and understanding," *npj Computational Materials*, vol. 8, no. 1, p. 250, 2022. DOI: [10.1038/s41524-022-00935-z](https://doi.org/10.1038/s41524-022-00935-z).
- [3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).

- [4] L. M. Ghiringhelli, C. Carbogno, S. Levchenko, *et al.*, "Towards efficient data exchange and sharing for big-data driven materials science: Metadata and data formats," *npj computational materials*, vol. 3, no. 1, p. 46, 2017. DOI: [10.1038/s41524-017-0048-5](https://doi.org/10.1038/s41524-017-0048-5).
- [5] L. M. Ghiringhelli, C. Baldauf, T. Bereau, *et al.*, "Shared metadata for data-centric materials science," *arXiv preprint arXiv:2205.14774*, 2022. DOI: [10.48550/arXiv.2205.14774](https://doi.org/10.48550/arXiv.2205.14774).
- [6] European Organization For Nuclear Research and OpenAIRE, *Zenodo*, en, 2013. DOI: [10.25495/7GXXK-RD71](https://doi.org/10.25495/7GXXK-RD71). [Online]. Available: <https://www.zenodo.org/>.
- [7] M. Könnecke, F. A. Akeroyd, H. J. Bernstein, *et al.*, "The nexus data format," *Journal of applied crystallography*, vol. 48, no. 1, pp. 301–305, 2015. DOI: [10.1107/S1600576714027575](https://doi.org/10.1107/S1600576714027575).
- [8] J. Starr, J. Ashton, A. Barton, *et al.*, *Datacite metadata schema for the publication and citation of research data, version 3*, 2013. DOI: [10.5438/0008](https://doi.org/10.5438/0008). [Online]. Available: https://schema.datacite.org/meta/kernel-3.0/doc/DataCite-MetadataKernel_v3.0.pdf.
- [9] R. Albertoni, D. Browning, S. Cox, A. G. Beltran, A. Perego, and P. Winstanley, *Data catalog vocabulary (dcat) - version 2*, 2020. [Online]. Available: <https://www.w3.org/TR/vocab-dcat-2/>.
- [10] C. W. Andersen, R. Armiento, E. Blokhin, *et al.*, "Optimade, an api for exchanging materials data," *Scientific data*, vol. 8, no. 1, pp. 1–10, 2021. DOI: [10.1038/s41597-021-00974-z](https://doi.org/10.1038/s41597-021-00974-z).
- [11] D. Miller, J. Whitlock, M. Gardiner, M. Ralphson, R. Ratovsky, and U. Sarid, *Openapi specification v3.0.3*, 2020. [Online]. Available: <https://spec.openapis.org/oas/v3.0.3>.
- [12] L. Himanen, P. Rinke, and A. S. Foster, "Materials structure genealogy and high-throughput topological classification of surfaces and 2d materials," *npj Computational Materials*, vol. 4, no. 1, pp. 1–10, 2018. DOI: [10.1038/s41524-018-0107-6](https://doi.org/10.1038/s41524-018-0107-6).
- [13] A. H. Larsen, J. J. Mortensen, J. Blomqvist, *et al.*, "The atomic simulation environment—a python library for working with atoms," *Journal of Physics: Condensed Matter*, vol. 29, no. 27, p. 273 002, 2017. DOI: [10.1088/1361-648X/aa680e](https://doi.org/10.1088/1361-648X/aa680e).
- [14] S. P. Ong, W. D. Richards, A. Jain, *et al.*, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," *Computational Materials Science*, vol. 68, pp. 314–319, 2013. DOI: [10.1016/j.commatsci.2012.10.028](https://doi.org/10.1016/j.commatsci.2012.10.028).
- [15] A. Jain, S. P. Ong, G. Hautier, *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL materials*, vol. 1, no. 1, p. 011 002, 2013. DOI: [10.1063/1.4812323](https://doi.org/10.1063/1.4812323).
- [16] S. Curtarolo, W. Setyawan, G. L. Hart, *et al.*, "Aflow: An automatic framework for high-throughput materials discovery," *Computational Materials Science*, vol. 58, pp. 218–226, 2012. DOI: [10.1016/j.commatsci.2012.02.005](https://doi.org/10.1016/j.commatsci.2012.02.005).
- [17] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd)," *Jom*, vol. 65, no. 11, pp. 1501–1509, 2013. DOI: [10.1007/s11837-013-0755-4](https://doi.org/10.1007/s11837-013-0755-4).
- [18] P. Ewels, T. Sikora, V. Serin, C. P. Ewels, and L. Lajaunie, "A complete overhaul of the electron energy-loss spectroscopy and x-ray absorption spectroscopy database: Eelsdb.eu," *Microscopy and Microanalysis*, vol. 22, pp. 717–724, Feb. 2016, ISSN: 1435-8115. DOI: [10.1017/S1431927616000179](https://doi.org/10.1017/S1431927616000179). [Online]. Available: http://journals.cambridge.org/article_S1431927616000179.

- [19] T. J. Jacobsson, A. Hultqvist, A. García-Fernández, *et al.*, "An open-access database and analysis tool for perovskite solar cells based on the fair data principles," *Nature Energy*, vol. 7, no. 1, pp. 107–115, Jan. 2022, ISSN: 2058-7546. DOI: [10.1038/s41560-021-00941-3](https://doi.org/10.1038/s41560-021-00941-3). [Online]. Available: <https://doi.org/10.1038/s41560-021-00941-3>.

Isn't a Number and a URL Enough?

Why PIDs Matter and Technical Solutions Alone are not Sufficient.

Stephanie Hagemann-Wilholt¹[\[https://orcid.org/0000-0002-0474-2410\]](https://orcid.org/0000-0002-0474-2410), Antonia Schrader²[\[https://orcid.org/0000-0001-7080-634X\]](https://orcid.org/0000-0001-7080-634X), and Andreas Czerniak³[\[https://orcid.org/0000-0003-3883-4169\]](https://orcid.org/0000-0003-3883-4169)

¹ Technische Informationsbibliothek (TIB), Germany

² Helmholtz Association, Germany

³ Bielefeld University Library, Germany

Abstract. Persistent identifiers (PIDs) are an integral part of research data management and can be found throughout the entire lifecycle of research data. However, their ability to function – to ensure persistence – depends on numerous factors: technical infrastructure, international standards and best practices and their dissemination, agreements on long-term governance of infrastructures, etc. Their applicability is diverse and requires adaptation to the resources and entities referenced by them. The paper describes two projects – PID4NFDI and PID Network Germany – that aim to address these challenges.

Keywords: Persistent Identifier, Research Data Management, Metadata, Base Services

1. FAIR Research Data Management – are PIDs an Additional Extra or Mandatory?

Imagine you get a dataset from a colleague to reuse for your own research. You know the name of the data producer, the data set has a working title and a versioning. The data producer also gives you some contextual information that is not in the dataset itself. This is all very helpful. But how do you cite this dataset? What if others would also like to find and reuse this dataset? Wouldn't it be nice to make the contextual information persistently available? These questions are already addressed by the endeavours of research data management (RDM), and its central efforts are to make data and its metadata findable, accessible, interoperable and re-usable. The backbone of all this effort is Persistent Identifiers (PIDs). The FAIR principles ask for the assignment of PIDs, the description of data with rich, standardised and machine-readable metadata and its long-term availability, referencing to other resources via PIDs, licensing and provenance information [1].

Therefore, PIDs have a significant impact on the entire life cycle of research data (RD) from the application for funding to the re-use of data and offer numerous advantages. They enable unique identification of research resources such as data, publications, or code; of persons, institutions, projects, grants etc [2]. PIDs support publication processes as well as discoverability and citability of research results. As widely accepted means to identify and link research entities, they are embedded in already established globally operating infrastructures and existing communities, which provide standards and best practices as a prerequisite for the harmonisation and thus the interoperability of RD metadata.

PIDs are an integral part of RDM services such as research data repositories, research information systems or knowledge graphs. A key aspect is that PIDs are established at the global level. They reduce administrative overhead, and the error rate of metadata ingestions and updates. PID metadata standards allow the linking of research entities and increase the global visibility and reusability of research resources as well as their accountability. Therefore, they are an important contribution to the quality assurance of research data and contribute to the reproducibility and reusability of research results.

Type and scope of entities described with PIDs in the context of RDM are continuously growing and gaining more and more maturity, but the landscape of actors, services, infrastructures and use cases is scattered. To leverage synergies within the existing structures, *PID4NFDI* and *PID Network Germany* aim to capture and analyse this landscape and to identify ways to consolidate it, to build and intensify networks and to establish international best practice.

2. PID4NFDI, PID Network Germany and their key objectives

PID4NFDI applies for funding within Base4NFDI to build a NFDI-wide PID service in line with the strategic goals of NFDI. It was initiated by the PID Working Group [3] in the section Common Infrastructures. DataCite, the Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG), the Helmholtz Open Science Office, and the German National Library of Science and Technology (TIB) are partners in the proposed project. Discipline-specific NFDI members will support the project as use case partners.

The project focuses on enhancing and optimising PID use in NFDI infrastructure services as a basic element of FAIR data management. To this end, technical, organisational, metadata, training and network aspects are tackled to serve the needs of existing and emerging services and communities as well as the NFDI as a whole. The aim is to harmonise requirements, to promote the integration of PIDs and to develop individual as well as central solutions and support for identified gaps. *PID4NFDI* aims to address these different use cases that are highly dependent on the type of research and nature of the data produced and used. Different research fields and methodologies each have their own requirements for the uses and benefits of PIDs: Which entities should be described? How granular are they? How long should they actually be available? Does the data have an impact on other research?

In the application phase, we already identified initial needs and requirements for an NFDI-wide PID infrastructure via a survey [4] and a stakeholder workshop [5]. Among other things, needs for an overview of PIDs and their scopes, the integration of subject-specific metadata, PIDs for entities such as instruments, ontologies, code, physical objects, projects, recommendations for dealing with ephemeral resources, granularity etc. have emerged.

PID Network Germany [6] is a DFG-funded project by DataCite, the German National Library (DNB), the Helmholtz Open Science Office, the German National Library of Science and Technology (TIB) and the Bielefeld University Library. The project started in March 2023 and aims to establish a network of stakeholders around the persistent identification of people, organisations, and resources in the field of digital communication in science and culture, which promotes the dissemination and connection of PID systems in Germany. The focus will also be on identifying needs and optimization potential for existing PID systems and on embedding them in international knowledge graphs. The project findings will lead to recommendations in the order to create a national PID roadmap for Germany.

The endeavour of *PID4NFDI* is in line with *PID Network Germany* and will make the German scientific landscape more robust, transparent and accessible. *PID Network Germany* pursues an overarching approach across the boundaries of individual scientific institutions and research disciplines. In this context, *PID4NFDI* represents an important building block for addressing the specific requirements of the NFDI. Constant and intensive exchange between the

projects will ensure that synergies are used and strengths are bundled, and that the activities are carried out in a target-oriented and at the same time target-group-specific manner.

Author contributions

The listed authors have prepared and written this extended abstract (role: Writing – original draft according to [CReDiT guidelines](#), Contributor Roles Taxonomy).

Competing interests

The authors declare that there are no competing interests.

Funding

PID Network Germany is funded by the German Research Foundation (DFG).

Acknowledgement

Thanks to Britta Dreyer, Lars G. Svensson and the members of the PID Working Group for helpful comments on the initial draft of this abstract.

References

- [1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>.
- [2] Ferguson, Christine, McEntrye, Jo, Bunakov, Vasily, Lambert, Simon, Sandt, Stephanie van der, Kotarski, Rachael, Stewart, Sarah, MacEwan, Andrew, Fenner, Martin, Cruse, Patricia, Horik, René van, Dohna, Tina, Koop-Jacobsen, Ketil, Schindler, Uwe, & McCafferty, Siobhan. (2018). D3.1 Survey of Current PID Services Landscape (Version 1). Zenodo. <https://doi.org/10.5281/zenodo.1324296>.
- [3] Bingert, Sven, Brase, Jan, Burger, Felix, Dreyer, Britta, Hagemann-Wilholt, Stephanie, Vierkant, Paul, & Wieder, Philipp. (2022). Concept for Setting up the Persistent Identifier Services Working Group in the NFDI Section "Common Infrastructures" (1.0). Zenodo. <https://doi.org/10.5281/zenodo.6507760>.
- [4] Hagemann-Wilholt, Stephanie. (2023, February 13). PID4NFDI: Survey on PID Practices. Main results. Zenodo. <https://doi.org/10.5281/zenodo.7635791>.
- [5] Schrader, Antonia C., Arend, Daniel, Bach, Janete, Elger, Kirsten, Göller, Sandra, Hagemann-Wilholt, Stephanie, Krahl, Rolf, Lange, Matthias, Linke, David, Mayer, Desiree, Mutschke, Peter, Reimer, Lorenz, Scheidgen, Markus, Selzer, Michael, & Wieder, Philipp. (2023). Workshop on PIDs within NFDI (Version 1). Internal Workshop on PIDs within NFDI, Online. Zenodo. <https://doi.org/10.5281/zenodo.7635905>.
- [6] Bertelmann, R., Buys, M., Kett, J., Pampel, H., Pieper, D., Scholze, F., Sens, I., Burger, F., Dreyer, B., Glagla-Dietz, S., Hagemann-Wilholt, S., Hartmann, S., Schrader, A. C., Schirrwagen, J., Summann, F., Vierkant, P. (2023): PID Network Deutschland. Netzwerk für die Förderung von persistenten Identifikatoren in Wissenschaft und Kultur, Potsdam : Helmholtz Open Science Office. <https://doi.org/10.48440/os.helmholtz.059>.

Leveraging Terminology Services for FAIR Semantic Data Integration across NFDI Domains

How to Integrate Terminology Services Into Other Service Applications

Roman Baum¹[\[https://orcid.org/0000-0001-5246-9351\]](https://orcid.org/0000-0001-5246-9351) and Oliver Koepler²[\[https://orcid.org/0000-0003-3385-4232\]](https://orcid.org/0000-0003-3385-4232)

¹ ZB MED - Information Centre for Life Sciences, Cologne, Germany

² TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany

Abstract. The National Research Data Infrastructure (NFDI) strives to develop FAIR research data and data services for major scientific disciplines, using terminologies as a key factor for semantic annotations and semantic interoperability of data. Several NFDI consortia provide domain-specific terminologies through Terminology services or registries, offering access, search capabilities, visualization, and downloads. Prioritizing user-friendly access, terminology services seamlessly integrate semantic concepts into applications, often operating in the background to enable smooth semantic annotation and data interoperability. We present exemplary fields of application from selected disciplines and how terminology services support semantic search, user experience, annotation workflows, terminology curation and design.

Keywords: Terminology Services, FAIR metadata

1. Introduction

The National Research Data Infrastructure (NFDI) aims to create FAIR research data and data services for all major scientific disciplines. Terminologies are a critical success factor for describing research data. Terminologies are applied to define concepts in metadata schemas to enable consistent and structured descriptions of data and their relationships. Therefore, we are observing an increasing use of the application of terminologies across the NFDI consortia, i.e. their discussion and harmonization in the working groups of the NFDI section (Meta)data, Terminologies and Provenance. To support researchers, data stewards, ontology experts and other services, several NFDI consortia such as DataPLANT [1], NFDI4Biodiversity [2], NFDI4Chem [3], NFDI4Earth [4], NFDI4Health [5], NFDI4Ing [6], or NFDI4Objects [7] provide domain-specific terminologies in respective domain-specific terminology services or registries.

Use cases vary across the disciplines, but all of the services should be as user-friendly as possible when it comes to accessing and selecting concepts. In such a user-friendly approach, semantic concepts should be automatically integrated into an application, rather than requiring time-consuming manual input. In fact, many application areas are not so obvious. Terminology services are often used in the background, embedded in other services.

A terminology service could be the backbone of a research data infrastructure. There are several application areas, such as 1) semantic search, 2) user experience enhancement, or 3) annotation services, where terminology services can provide support.

2 Existing Terminology Services in the NFDI

Due to constraints of this format, we cannot fully describe all terminology services in all domains. Similar fields of application and approaches are valid for other services as well. We will focus on the two services SemLookP and the NFDI4Chem Terminology Service which are both based on the Ontology Lookup Service (OLS) [8].

2.1 SemLookP

SemLookP is a semantic lookup platform that provides terminologies from the fields of medicine, nutrition and life sciences. The access to these terminologies is realized via a graphical user interface (GUI) or via APIs. The GUI of SemLookP consists of several combined JavaScript based widgets. Such a widget uses the OLS API and combines the data received from the API with specific HTML components. These widgets can also be directly integrated into other applications.

2.2 NFDI4Chem Terminology Service

The NFDI4Chem Terminology Service provides access to a collection of ontologies relevant to the chemistry community. The collection was derived from an evaluation process [9] and is part of the Ontologies4Chem endeavor [10]. The Terminology Service provides faceted search inside ontologies, tree and lists views of concepts, properties and individuals and how to access them. The service not only provides an overview of ontologies in the domain, but also aims to support comparison and analysis across multiple ontologies for curation tasks. For this It offers an unified view of issues from the original ontology repositories within the Terminology GUI. The NFDI4Chem Terminology Service can also index and display SKOS vocabularies. The terminology service also provides a comprehensive API to retrieve all data and information and embed it into other services and applications.

3. Applications and Integration of Terminology Services

In the following we demonstrate some application and integration cases of terminology services in other applications mostly based on SemLookP and the NFDI4Chem Terminology Service. We will present 3 different topics where and how a terminology service could be easily integrated into an application.

3.1 Semantic Search

preVIEW is the first application we will focus on in this section. preVIEW is a user-friendly COVID19-related preprint viewer with advanced semantic search functionality. [11, 12] Since the terminologies from SemLookP are used for the annotation process, it is possible to use these concepts in the semantic search of preVIEW (see Figure 1).

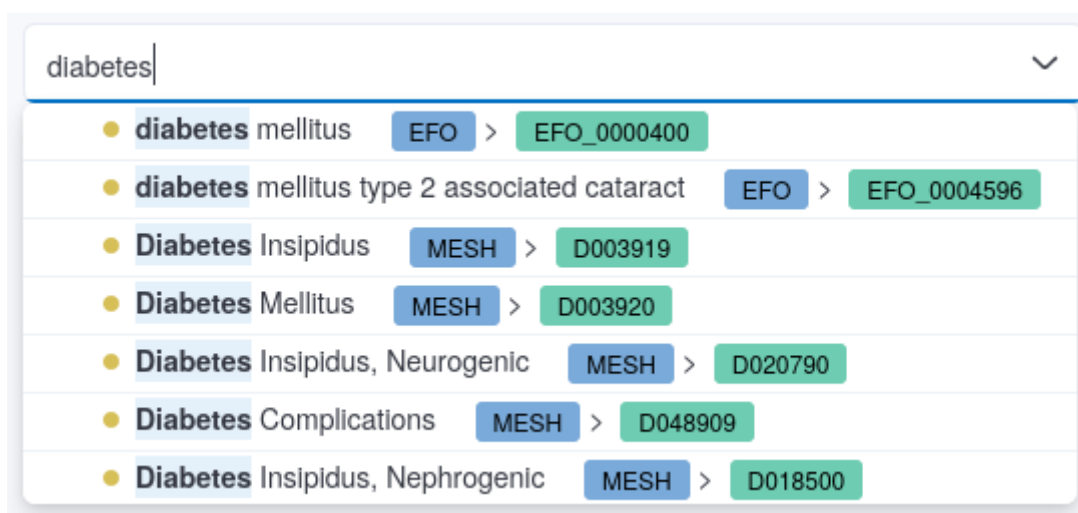


Figure 1. Autocomplete widget which is integrated into preVIEW.

There is also the GFBio Data Portal application. The GFBio Data Portal uses a RESTful API of the GFBio Terminology Service [13] for the annotation and the semantic search. Since the GFBio services are self-developed, it also proves that the integration of a terminology service into another application is possible with systems other than OLS.

3.2 User Experience Enhancement

To enhance the user experience, we will again focus on preVIEW. The annotations in preVIEW are highlighted with a background color. Clicking on an annotated concept opens a window. This window contains the metadata of the annotated concept (see Figure 2).

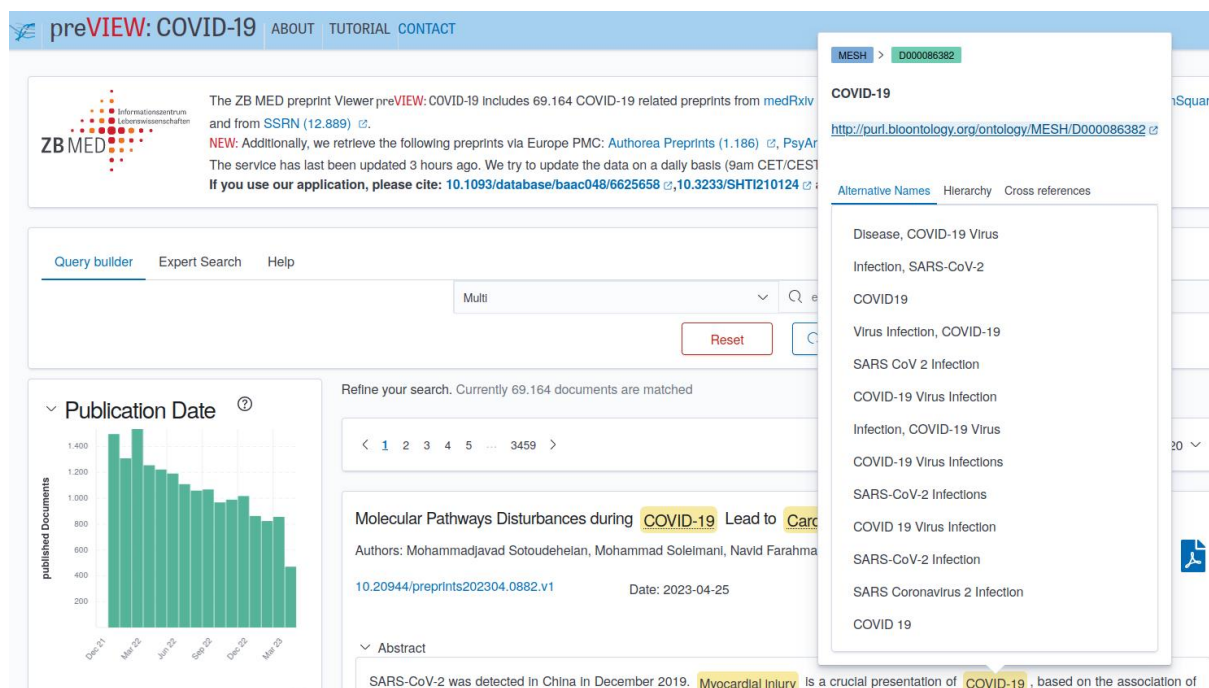


Figure 2. Metadata widget which is integrated in preVIEW.

Next to directly integrate widgets of a terminology service to another application it is also possible to receive the presented data via the API of a terminology service. This data could be used to develop custom HTML components. An example would be a semantic search filter. Imagine a user enters a specific term in a search bar. A hierarchical filter listing sub-

concepts could then appear next to the list of results. Such a sub-concept could act as a shortcut to further specify the search term.

3.3 Data Annotation Services

Many use cases of the NFDI4Chem Terminology Service can be described as service-to-service applications. The extensive API of the terminology service can be used by other services like the chemotion electronic lab notebook (ELN) or data repositories like RADAR4Chem, nmrXiv or others. The ELN reuses concepts from the named reaction ontology or chemical methods ontology to annotate reactions in experiments or analytical methods applied for the generation of data.

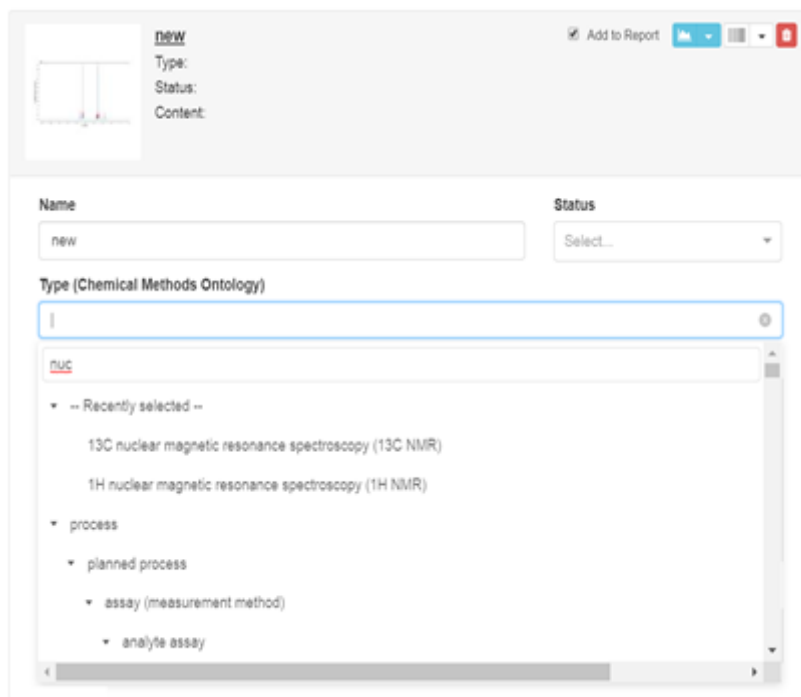


Figure 3. Reuse of CHMO terms in Chemotion ELN.

If not already provided by ELNs, repositories apply terms from ontologies for molecular entities, analytical methods, or processes to semantically annotate research data in their forms during the upload workflow.

This approach is freely transferable to other domains. For example, the NFDI4Health Metadata Annotation Workbench uses the API of SemLookP.

4. Conclusion and Outlook

In this work, we presented a very simple and smart solution for integrating terminology services into other applications by using the API or widgets.

However, domain-specific terminology services have limitations. These terminology services are isolated in their own domain. To solve this problem, terminology services could be connected and harmonized through an additional gateway in the future (see Figure 4). Such an approach would bridge the siloed solution and enable cross-domain fields of application.

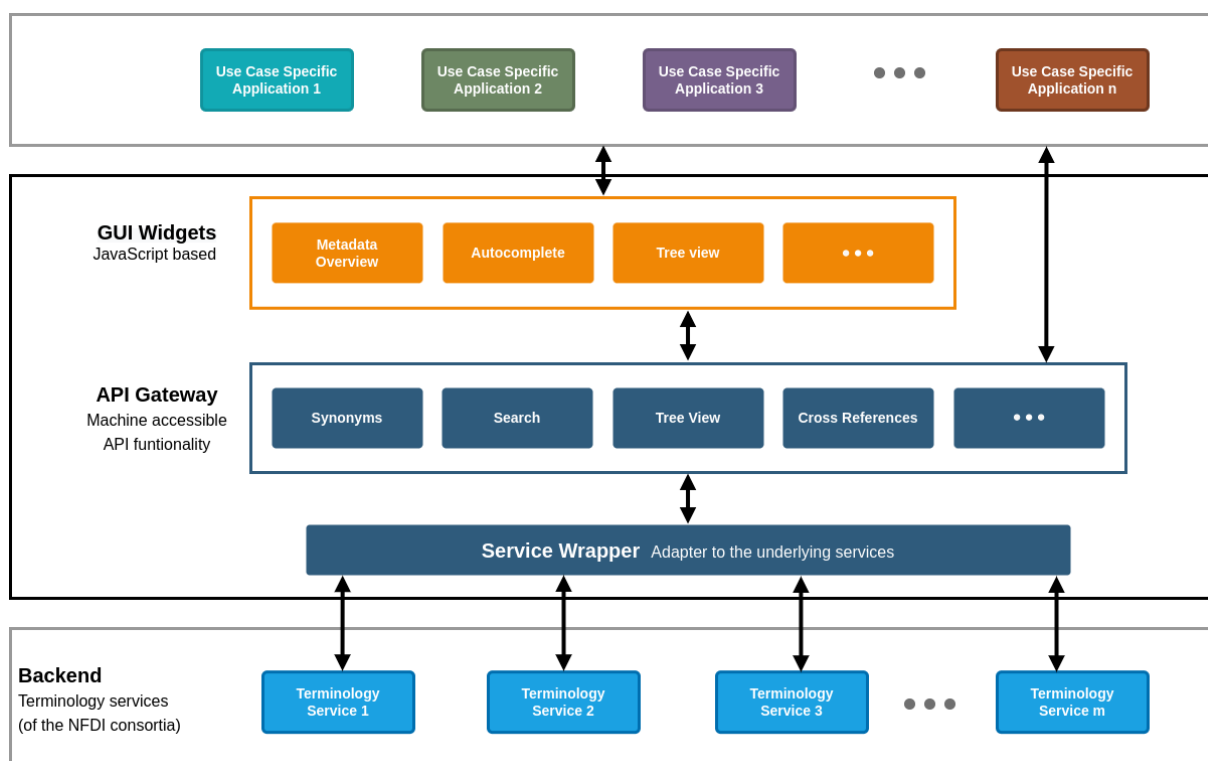


Figure 4. Architecture of a shared terminology API/widget tool set.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was done as part of the Consortia NFDI4Health and NFDI4Chem. We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - grant numbers: 442326535 (NFDI4Health), 441958208 (NFDI4Chem).

References

1. SwateOntology DB / Swobup, <https://nfdi4plants.org/content/service.html> (last accessed 25 April 2023)
2. GFBio Terminology Service, <https://terminologies.gfbio.org/> (last accessed 25 April 2023)
3. NFDI4Chem Terminology Service, <https://terminology.nfdi4chem.de> (last accessed 25 April 2023)
4. GEMET - GEneral Multilingual Environmental Thesaurus, <https://www.eionet.europa.eu/gemet/> (last accessed 25 April 2023)
5. SemLookP, <https://semanticlookup.zbmed.de/> (last accessed 25 April 2023)
6. NFDI4Ing Terminology Service, <https://terminology.nfdi4ing.de/ts/> (last accessed 25 April 2023)
7. DANTE, <https://api.dante.gbv.de/> (last accessed 25 April 2023)
8. S. Jupp, T. Burdett, C. Leroy, and H. Parkinson, "A new Ontology Lookup Service at EMBL-EBI", presented at the Workshop on Semantic Web Applications and Tools for Life Sciences, 2015. Accessed: Apr. 25, 2023. [Online]. Available:

- <https://www.semanticscholar.org/paper/A-new-Ontology-Lookup-Service-at-EMBL-EBI-Jupp-Burdett/b83bfbfc1f2f08e5b88af5ef65ef2a8687ac4112>
9. P. Strömert, J. Hunold, A. Castro, S. Neumann, and O. Koepler, "Ontologies4Chem: the landscape of ontologies in chemistry," *Pure Appl. Chem.*, Mar. 2022, doi: <https://doi.org/10.1515/pac-2021-2007>.
 10. P. Strömert, J. Hunold, and O. Koepler, "1st Ontologies4Chem Workshop – Ontologies for chemistry," Sep. 07, 2022, doi: <https://doi.org/10.25798/frmp-sn04>.
 11. L. Langnickel, R. Baum, J. Darms, S. Madan, and J. Fluck "COVID-19 preVIEW: Semantic Search to Explore COVID-19 Research Preprints," *Public Health and Informatics*, vol.281, pp. 78-82, May, 2021, doi: <https://doi.org/10.3233/SHTI210124>
 12. L. Langnickel, J. Darms, R. Baum, and J. Fluck "preVIEW: from a fast prototype towards a sustainable semantic search system for central access to COVID-19 preprints," *EAHIL*, vol.17, no.3, pp. 8-14, Sep., 2021, doi: <https://doi.org/10.32384/jeahil17484>
 13. N. Karam, C. Müller-Birn, M. Gleisberg et al., "A Terminology Service Supporting Semantic Annotation, Integration, Discovery and Analysis of Interdisciplinary Research Data," *Datenbank Spektrum* 16, pp 195–205, Oct. 05, 2016, doi: <https://doi.org/10.1007/s13222-016-0231-8>

Finding a Common Ground for NFDI Terminologies

Proposing I-ADOPT as a NFDI Wide Semantic Layer

Robert Huber¹, Naouel Karam^{2,3}[\[https://orcid.org/0000-0002-6762-6417\]](https://orcid.org/0000-0002-6762-6417), Oliver Koepler⁴[\[https://orcid.org/0000-0003-3385-4232\]](https://orcid.org/0000-0003-3385-4232), and Philip Strömert⁴[\[https://orcid.org/0000-0002-1595-3213\]](https://orcid.org/0000-0002-1595-3213)

¹ University of Bremen

² Institute for Applied Informatics (InfAI), Leipzig, Germany

³ Fraunhofer FOKUS , Berlin, Germany

⁴ TIB - Leibniz Information Centre for Science and Technology

Terminology Harmonisation in NFDI

Top-level ontologies (TLOs) and mid-level ontologies (MLOs) play a very important role in enabling semantic interoperability between domain-specific ontologies by providing a general structure and common high level entities and relationships for classifying and interlinking domain-specific concepts. A number of such ontologies have been proposed for different purposes. Unfortunately, due to different ontology design patterns, some of these ontologies are not interoperable out of the box. In order to increase the cross-domain interoperability of research data within NFDI and the EOSC, we need to harmonise the used TLO and MLO concepts to a common ground. The Section-Metadata working group Ontology Harmonisation & Mapping was formed to coordinate and guide such an alignment work, by recommending, providing and/or developing mappings, frameworks and tools [1]. We started to analyse which ontologies are used among many NFDI consortia and found that using only one specific TLO & MLO framework will not meet the different ontological requirements. Consequently, we need to provide formal mappings between many commonly used concepts, such as process, information, characteristic or method, defined in a variety of common TLO and MLO as well as SKOS vocabularies and other reference terminologies. Since this will be a complex and labour intensive process, we must also look at less complex solutions for interdisciplinary data integration. One approach could be to focus first on the observational part of research data.

I-ADOPT as a Possible NFDI Wide Interoperable Variables Description Framework

Observational data play a crucial role in many scientific disciplines and the need for combined analysis of findings across disciplines is increasing. Although containing similar information from the semantic point of view, observations are described using a variety of metadata standards and semantic resources, introducing a high level of heterogeneity between data management systems. In order to cope with those issues, members of the Working Group Interoperable Descriptions of Observable Property Terminology (RDA I-ADOPT WG) proposed an interoperability framework and an ontology to semantically describe observable properties [2]. The I-ADOPT ontology describes a variable (i.e. what is observed, measured, or derived) as a

complex concept consisting of at least an entity of interest and its property. Additional components can be used to describe the context as well as constraints on the entities. It is possible to use this framework in a broader context with other representations of observations (e.g. *Characteristic* in OBOE or *study design* in OBI).

In Figure 1, we show the implementation of the framework on a complex variable example “Egg production rate of female *Calanus finmarchicus*”. The property “rate” is applied to the object of interest “egg production” in the context of the entity “*Calanus finmarchicus*”. The components of such an I-ADOPT based variable are meant to be mapped to other commonly used terminologies [3], as for example, the Phenotype And Trait Ontology (PATO) to describe properties and constraints. Integrating other terminologies this way, adds a knowledge layer to terminologies that lack such expressivity or can reduce the need for additional instances when more general modelling patterns are used (see Figure 1c). An overall mapping effort reduction is expected, as only the variable components reference terms need to be mapped for interdisciplinary data integration.

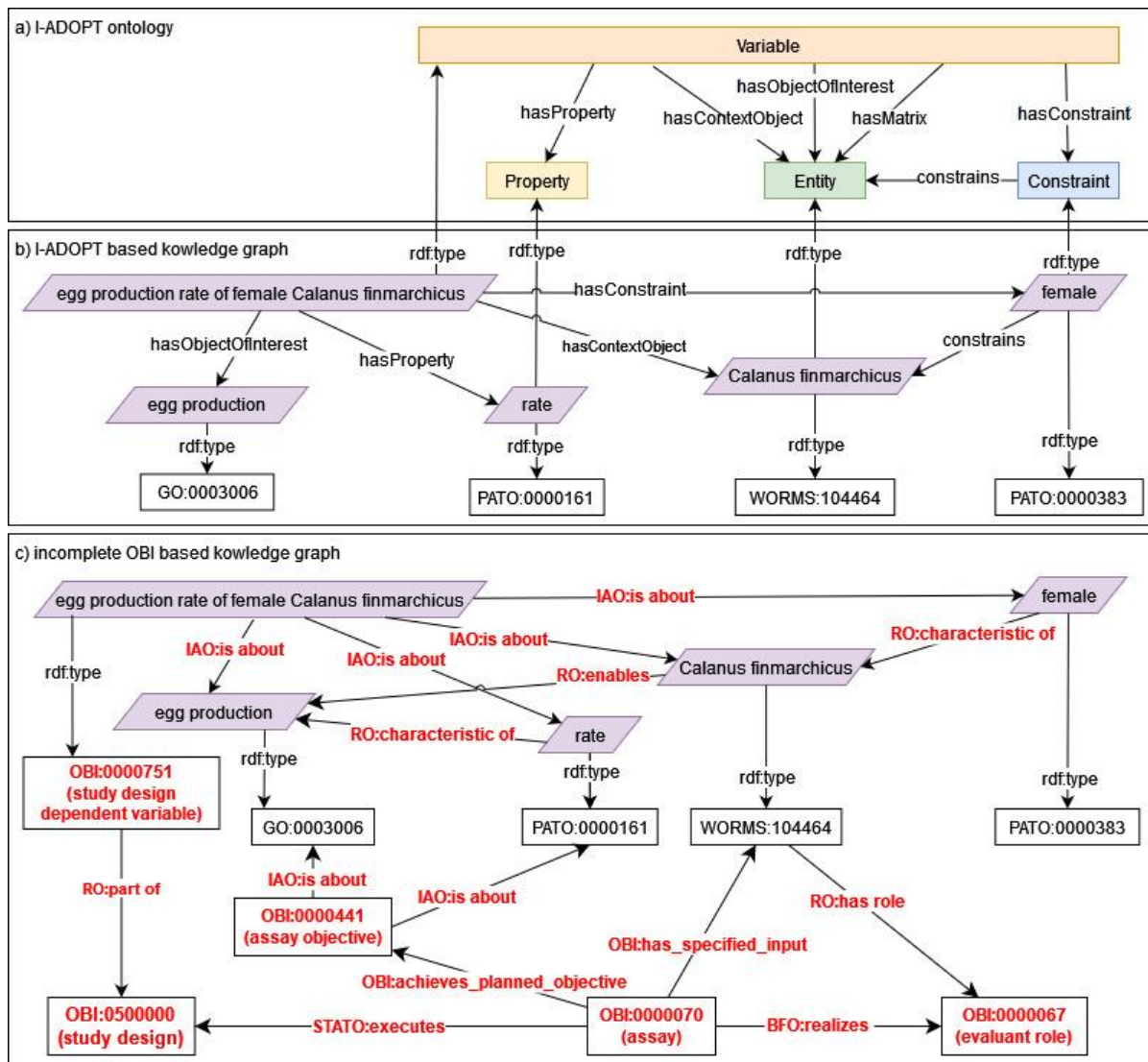


Figure 1: Semantic differences between the I-ADOPT and OBO Foundry framework

In the context of NFDI4Biodiversity [4], we are implementing the I-ADOPT framework to enable interoperability between what is observed or measured in different contexts in order to facilitate search and data integration between multiple data provider partners like among others PAN-GAEA [5]. The semantic model of Schema.org’s *Dataset* type is connected through the *variableMeasured* property to the I-ADOPT model. All necessary terminologies for the I-ADOPT

components are collected in our terminology repository and service [6] together with their mappings [7].

I-ADOPT Annotator

A widespread adoption of such a framework depends on the automation of variable annotation. Yet, mapping I-ADOPT's variable components to external terminologies is not a trivial task. It is especially difficult with ontologies that cover a much broader domain and use a more complex axiomatisation, as concepts of such ontologies might possibly be mapped to I-ADOPT's *entity* as well as *property* or *constraint* concept. This makes it hard to decide automatically with high confidence which mapping fits the actual context, as it requires more complex algorithms. For highly specialised taxonomies, like WORMS (the World Register of Marine Species) [8] or ITIS (the Integrated Taxonomic Information System) [9], on the other hand, an automatic and high confidence based mapping to I-ADOPT is rather simple.

For PANGAEA, we have made first efforts to map the assignment of terms in its parameters [10] to I-ADOPT. To this end, an existing annotation service [11] is currently extended to use rules, ontologies, and grammar to decide how to map a term recognized in a parameter to I-ADOPT. Figure 2 shows a possible output of the service as JSON. In the example, the term 'Calanus finmarchicus' is recognized as 'ContextObject' because all WORMS entries are considered to be of type *entity* and since the precise I-ADOPT role (e.g. object of interest, matrix or context object) cannot be distinguished the most generic role 'ContextObject' is listed here, 'egg production rate' is recognised as a *property* based on the identification of the term *rate* which is listed in PATO as a direct child of a term that is a subclass of *quality* and in contrast the PATO term *female* is considered as a *constraint* since it has a larger graph distance to a PATO entry which is a subclass of a quantity.


```

{
  "parameter": "Calanus finmarchicus, egg production rate per female, wet",
  "dim_match": {},
  "text_match": {
    {
      "fragment": "Calanus finmarchicus",
      "start_offset": 0,
      "end_offset": 20,
      "term": {
        {
          "id": 1053596,
          "name": "Calanus finmarchicus",
          "score": 103.663475,
          "terminology": "WoRMS, Aphia 1.0",
          "description_uri": "https://www.marinespecies.org/aphia.php?p=taxdetails&id=104464",
          "iadopt_type": "ContextObject"
        }
      }
    },
    {
      "fragment": "egg production rate",
      "start_offset": 22,
      "end_offset": 41,
      "term": {
        {
          "id": 1073136,
          "name": "rate",
          "score": 46.896287,
          "terminology": "PATO",
          "description_uri": "http://purl.obolibrary.org/obo/PATO_0000161",
          "iadopt_type": "Property"
        }
      }
    },
    {
      "fragment": "female",
      "start_offset": 46,
      "end_offset": 52,
      "term": {
        {
          "id": 1074133,
          "name": "female",
          "score": 46.831738,
          "terminology": "PATO",
          "description_uri": "http://purl.obolibrary.org/obo/PATO_0000383",
          "iadopt_type": "Constraint"
        }
      }
    }
  }
}

```

Figure 2: JSON output of Param-Annotator

Conclusion

Our aim within the NFDI working group and related international initiatives is to provide both a high level linking of the NFDI ecosystem terminologies through the mapping of TLOs and MLOs as well as concentrate our efforts on specific ones which could rapidly lead to interoperability across disciplines.

Competing interests

The authors declare that they have no competing interests.

Funding

Part of this work has been done within the NFDI4Chem and NFDI4Biodiversity project, both funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the NFDI Funding Programme (grant no. 441958208 & 442032008).

Acknowledgement

We want to thank all participants of the Section Metadata Ontology Harmonisation & Mapping Working Group for their feedback and input to this work.

References

1. I. Anders, T. Arera-Rütenik, S. Arndt, R. Baum, N. Betancort, I. Blümel, C. Busse, A. Daniel, F. Engel, L. Ghiringelli, S. Hachinger, H. Israel, N. Karam, A. Kranz, R. Lenz, D. Linke, T. Petrenko, L. Rossenova, D. Schulz, C. Weiland, H. Weiland, C. Wiljes, M. Özaslan, N. Kockmann. (2022). Ontology Harmonization and Mapping - Working Group Charter (NFDI section-metadata) (1.0). Zenodo. <https://doi.org/10.5281/zenodo.6726519>
2. B. Magagna, I. Rosati, M. Stoica, S. Schindler, G. Moncoiffe, A. Devaraju, J. Peterseil, R. Huber: The I-ADOPT Interoperability Framework for FAIRer Data Descriptions of Biodiversity. JOWO 2021.
3. S. Schindler, G. Moncoiffé. (2022). List of Terminologies. <https://i-adopt.github.io/terminologies/list/all>. Accessed 2023-04-26.
4. GFBio e.V.. (2023). NFDI4Biodiversity Homepage. <https://nfdi4biodiversity.org>. Accessed 2023-04-26.
5. PANGAEA, Data Publisher for Earth & Environmental Science. (2023). <https://pangaea.de>. Accessed 2023-04-26.
6. Karam, N.; Fichtmueller, D.; Gleisberg, M.; Becker, F.; Tolksdorf, R.; Müller-Birn, C.; Paschke, A. & Güntsch, A. (eds.) 2014 onwards. The Terminology Service of the German Federation for Biological Data (GFBio) - Service of semantic technologies in scientific environments. - <http://terminologies.gfbio.org>. Accessed 2023-04-26.
7. N. Karam, A. Khiat, A. Algergawy, M. Sattler, C. Weiland, M. Schmidt: Matching biodiversity and ecology ontologies: challenges and evaluation results. Knowl. Eng. Rev. 35: e9 (2020).
8. WoRMS Editorial Board (2023). World Register of Marine Species. Available from <https://www.marinespecies.org> at VLIZ. Accessed 2023-04-26. <https://doi.org/10.14284/170>
9. Retrieved [April, 26, 2023], from the Integrated Taxonomic Information System (ITIS), www.itis.gov. <https://doi.org/10.5066/F7KH0KBK>
10. M. Diepenbroek et al (2017): Terminology supported archiving and publication of environmental science data in PANGAEA, Journal of Biotechnology 261 (2017) 177-186, <https://doi.org/10.1016/j.jbiotec.2017.07.016>
11. PANGAEA Web Services Server. (2023). ws.pangaea.de/param-annotator. Accessed 2023-04-26.

A survey on the current status of Research Data Management in Mecklenburg-Vorpommern

Preliminary results for a questionnaire study among researchers

Max Schröder^{1,2}[\[https://orcid.org/0000-0003-1522-494X\]](https://orcid.org/0000-0003-1522-494X),
Sascha Genehr^{1,2}[\[https://orcid.org/0000-0002-1702-6878\]](https://orcid.org/0000-0002-1702-6878),
Rüdiger Köhling³[\[https://orcid.org/0000-0003-3330-4898\]](https://orcid.org/0000-0003-3330-4898),
Stefan Schmidt⁴[\[https://orcid.org/0000-0003-2450-9795\]](https://orcid.org/0000-0003-2450-9795),
Ralf Schneider⁵[\[https://orcid.org/0000-0002-4492-8869\]](https://orcid.org/0000-0002-4492-8869),
Sascha Spors²[\[https://orcid.org/0000-0001-7225-9992\]](https://orcid.org/0000-0001-7225-9992),
Gero Szepannek⁶[\[https://orcid.org/0000-0001-8456-1283\]](https://orcid.org/0000-0001-8456-1283),
Dagmar Waltemath⁷[\[https://orcid.org/0000-0002-5886-5563\]](https://orcid.org/0000-0002-5886-5563), and
Frank Krüger⁸[\[https://orcid.org/0000-0002-7925-3363\]](https://orcid.org/0000-0002-7925-3363)

¹University Library, University of Rostock, Germany

²Institute of Communications Engineering, University of Rostock, Germany

³Oscar Langendorff Institute for Physiology, University Medical Center Rostock, Germany

⁴Department of Health, Nursing, Management, Neubrandenburg University of Applied Sciences, Germany

⁵Institute of Physics, University of Greifswald, Germany

⁶School of Business Studies, Stralsund University of Applied Sciences, Germany

⁷Medical Informatics Laboratory, University Medicine Greifswald, Germany

⁸Department of Electrical Engineering and Computer Science, Wismar University of Applied Sciences, Germany

Abstract: High quality research data management (RDM) is essential to support state of the art comprehensible and reproducible research processes and, thus, foster the sustainable production of novel and trustworthy research findings. While there are lots of national and international initiatives supporting researchers in all respects of RDM, the local infrastructures provide the foundation for these concepts. In this contribution, we present preliminary results of a study that collects the requirements of researchers on these local infrastructures in all seven higher education research institutions in Mecklenburg-Vorpommern (MV).

Keywords: Research Data Management, Mecklenburg-Vorpommern, Mecklenburg-Western Pomerania, Higher Education Research, Online Survey

The initial development of the study was driven by the goal of measuring the needs of the federated state MV in terms of RDM. For this purpose, an online questionnaire was developed that aims at collecting information about

1. What is the current state of RDM in MV?
2. Which aspects influence the implementation of RDM?

Table 1. The number of survey participants for each institution.

Institution	# Participants
University of Greifswald	40
University of Rostock	161
University Medical Center Greifswald	48
University Medical Center Rostock	41
Neubrandenburg University of Applied Sciences	3
Stralsund University of Applied Sciences	12
Wismar University of Applied Sciences	23
Sum	328

3. Are there discipline-specific differences, and what role does the experience and the position of researchers play in the implementation of RDM?
4. How much effort do researchers put into specific aspects of RDM and how do they rate their significance?

By conducting this study, we aim at deriving future directions of institutional support in MV. Thus, this study can provide the foundation for a strategy of the federated state MV with respect to RDM.

Similar studies have been conducted at institutional level, e.g. Arndt, Glatz, Hummel, *et al.* have investigated how researchers deal with their data at the University of Applied Sciences Potsdam [1], and also discipline-specific level, e.g. Senft, Stahl, and Svoboda have surveyed RDM in the agricultural sciences [2]. While we considered these studies during the development of the questionnaire, the focus of our study is different so that most questions are not comparable to the other investigations. However, some questions that fit into the focus have been integrated from [1], [2].

To obtain information for our goal, the questionnaire contained eight content sections with an additional section at the beginning in order to ask for the participants' informed consent. In order to ease answering, the questionnaire was conducted in German language. The eight content sections are oriented to the research data life cycle [3]:

1. **General questions** that ask for the three most important aspects of RDM in the participants' research activity, as well as several summary questions on participants' RDM practice such as: "Have you ever published research data?". While aiming at introducing participants into the spectrum of topics in RDM, further sections, that do not apply for the participant, will be hidden based on selections.
2. **Research data management in the working group** contains questions about participants' basic RDM practice such as the methods they typically use to collect their research data, the most important data formats, and the amount of data they handle.
3. **Concepts, methods, technologies and tools for research data management** contains questions about the usage frequency of several state-of-the-art RDM concepts etc., e.g. data management plans, minimal information standards, ORCID, or electronic lab notebooks (ELN).
4. **Data storage and sharing** asks participants about the frequency of different storage systems and sharing methods, e.g. mobile storage media, institutional and external repositories, and institutional and external cloud systems. Furthermore, questions regarding the effort and the significance of sharing with different stake-

holders, e.g. the working group and the project partners, other researchers, or reviewers, are asked.

5. **Archiving and publishing** concerns different media and platforms for archival and publication, e.g. institutional and external repositories, as well as questions regarding the kind of publication, e.g. as supplement material or in a data repository, and the type of data, e.g. raw data or cleaned data. Furthermore, this section also contains questions regarding the effort and significance for the participants' research.
6. **Re-use** collects information about sources used to search for existing data and from which data has actually been re-used, as well as the source of data availability inquiries towards the participants. Again, the effort and significance of these aspects has been surveyed in advance.
7. **Standards and processes** asks participants about several aspects regarding their research processes, e.g. which standards they use as well as the source of these standards. In addition, the significance of several technical aspects for the participants' research is requested, e.g. the storage of data within the institution, the availability of a graphical user interface (GUI) and the availability of an application programming interface (API) for tools and services. Participants are also asked to rate the effort and significance of some standard research tasks, e.g. textual documentation of research investigations independent of a publication.
8. **Academic background** contains questions regarding the highest academic degree, the research discipline according to the DFG, the role of the person in the working group, how the participants rate their knowledge, the probability of attending a training, and the importance of RDM for their research.

After a pretest, the revised questionnaire was employed to collect answers from all higher education research institutions in MV using the evasys¹ survey platform at the University of Rostock. The institutions that have been surveyed are:

1. University of Greifswald,
2. University of Rostock,
3. University Medical Center Greifswald,
4. University Medical Center Rostock,
5. Neubrandenburg University of Applied Sciences,
6. Stralsund University of Applied Sciences, and
7. Wismar University of Applied Sciences.

Universities of public administration or labour studies and private institutions are not included in the study in order to reflect the target group: researchers in higher education research institutions of the federated state MV. The invitations for the questionnaire were sent to professors, scientific staff, and doctoral candidates in every institution via e-mail utilizing corresponding mailing lists. Two weeks after the initial invitation, in most institutions, a reminder was sent. The questionnaire was open for four weeks at each institution and the data collection was performed between March 13th, 2023 and April 30th, 2023.

In total, 328 participants filled the survey. To the best of our knowledge, this is the largest survey on RDM in Germany and MV in particular. The participants' distribution across the institutions is presented in Table 1. Figure 1 presents the answers of the participants regarding their role (top) and their primary research discipline according to the DFG subject classification (bottom). Interestingly, all 14 disciplines are represented

¹<https://evasys.de/evasys/>

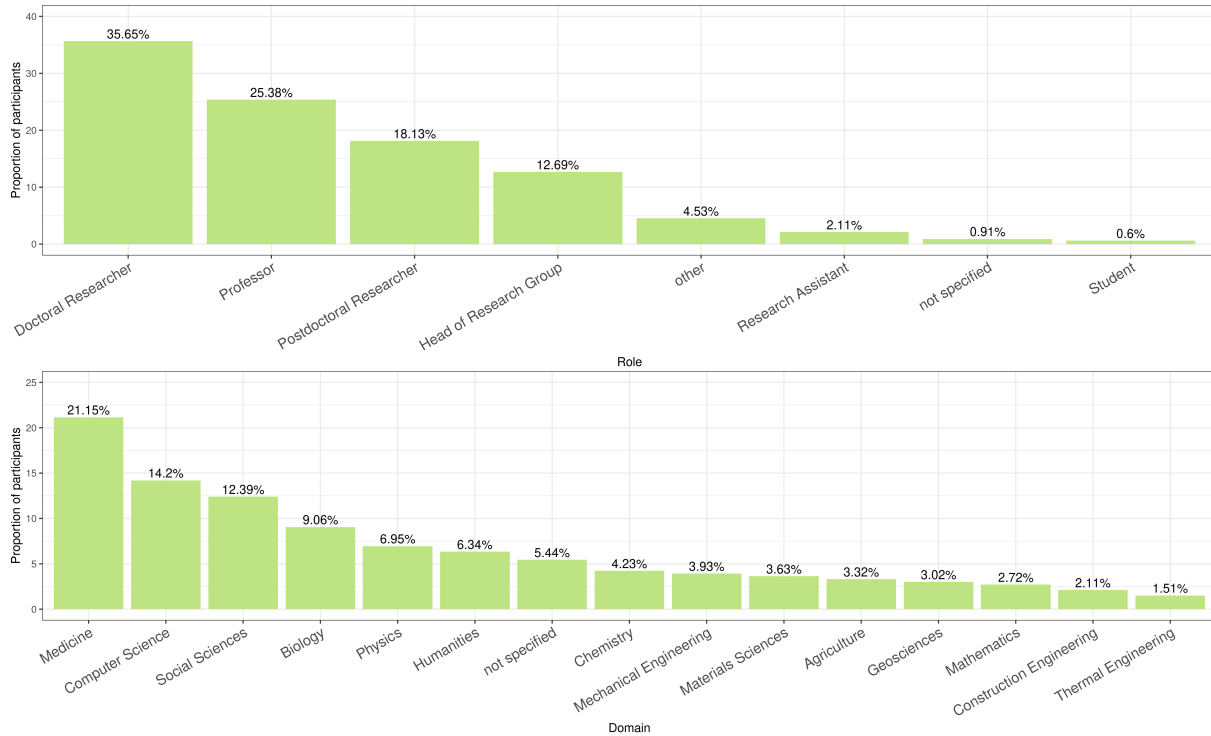


Figure 1. Distribution of participants' disciplines and roles aggregated for all participating institutions. Please note that discipline names are shortened to improve the visualization.

within the participants though the amount is different between the disciplines. In this contribution we will present preliminary results of this study.

Data availability statement

This article reports on the preliminary results of a survey regarding research data management across Mecklenburg-Vorpommern. The questionnaire is published at Zenodo: [4]. All data will be published together with a detailed analysis of the results of the survey.

Author contributions

MS: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Validation, Writing – original draft; **SG:** Writing – original draft; **RK:** Resources; **SSc:** Resources; **RS:** Resources; **SSp:** Funding acquisition, Supervision; **GS:** Resources; **DW:** Resources; **FK:** Conceptualization, Formal Analysis, Funding acquisition, Supervision, Visualization, Writing – original draft

Competing interests

The authors declare that they have no competing interests.

Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1270/2 – 299150580.

Acknowledgements

We highly acknowledge the valuable feedback of Prof. Dr. Gabriele Doblhammer, Maximilian Behnert-Brodhun, Meike Bielfeldt, Kai Budde, Anja Eggert, Milan Kostresevic, and Susanne Stähle during the development of the questionnaire. We also thank the team RDM of the Rostock University Library for their feedback on the questionnaire. Furthermore, we thank the Staff Office for University and Quality Development of the University of Rostock for the support during the sending of the invitations.

References

- [1] O. Arndt, L. Glatz, B. Hummel, M. Porst, W. Schabalowski, and S. Skubatz, *Umfrage zum Forschungsdatenmanagement an der FH Potsdam*, H. Neuroth and M. Ortgiese, Eds. Verlag der Fachhochschule Potsdam, p. 190, ISBN: 9783934329959.
- [2] M. Senft, U. Stahl, and N. Svoboda, "Research data management in agricultural sciences in germany: We are not yet where we want to be," *PLOS ONE*, vol. 17, no. 9, C. Pulvento, Ed., e0274677, Sep. 2022. DOI: [10.1371/journal.pone.0274677](https://doi.org/10.1371/journal.pone.0274677).
- [3] H. Enke, N. Fiedler, T. Fischer, *et al.*, *Leitfaden zum Forschungsdaten-Management: Handreichungen aus dem WissGrid-Projekt*, J. Ludwig and H. Enke, Eds. Glückstadt: Verlag Werner Hülsbusch, 2013, ISBN: 978-3-86488-032-2. [Online]. Available: <https://resolver.sub.uni-goettingen.de/purl?gro-2/14366>.
- [4] M. Schröder, S. Spors, and F. Krüger, *Questionnaire on the current status of research data management in mecklenburg-vorpommern*, de, 2023. DOI: [10.5281/ZENODO.8099171](https://doi.org/10.5281/ZENODO.8099171).

Machine-Actionable Metadata for Software and Software Management Plans for NFDI

Olga Giraldo¹[\[https://orcid.org/0000-0003-2978-8922\]](https://orcid.org/0000-0003-2978-8922), Danilo Dessi²[\[https://orcid.org/0000-0003-3843-3285\]](https://orcid.org/0000-0003-3843-3285), Stefan Dietze²[\[https://orcid.org/0009-0001-4364-9243\]](https://orcid.org/0009-0001-4364-9243), Dietrich Rebholz-Schuhmann^{1,3}[\[https://orcid.org/0000-0002-1018-0370\]](https://orcid.org/0000-0002-1018-0370), and Leyla Jael Castro¹[\[https://orcid.org/0000-0003-3986-0510\]](https://orcid.org/0000-0003-3986-0510)

¹ ZB MED Information Centre for Life Sciences, Cologne, Germany

² GESIS Leibniz Institute for Social Sciences, Cologne, Germany

³ University of Cologne, Cologne, Germany

Abstract. Research data is on its way to be recognized as a first-class citizen in research; however, and despite its importance for science, software still has a long way to go. Recent initiatives are paving the way, including FAIR for Research Software and Software Management Plans. A step further towards machine-actionability is adding a structured metadata layer. Here we discuss some metadata elements useful to represent software and integrate it into management plans, and how it could be of benefit for NFDI.

Keywords: Research Software, Management Plan, Metadata, Machine-Actionable

1. Background

Traditionally, research outcomes have been published in text-based scholarly publications, where data and software used (or produced) are (sometimes) briefly discussed. Rich metadata exists for scholarly publication, making it easier to extract data and use it to create insights and knowledge out of it, for instance co-citation or co-author networks. Combined with Natural Language Processing techniques, in particular text-mining and text-based embeddings, further analysis becomes possible. Data and software are nowadays recognized as key players for the advance of science; however, they are not yet first-class citizens when it comes to publication and citation.

The Findable, Accessible, Interoperable and Reusable (FAIR) guiding principles for data [1] favor the use of machine-actionable metadata, i.e., metadata semantically structured facilitating search and retrieval while also facilitating (semi)automatic integration and validation. FAIR principles have also boosted research data publication and citation. Although lagging behind, research software is also moving forward in this direction, one of the reasons being its importance in science reproducibility. Some efforts working to make software a first-class citizen in research are the community-endorsed FAIR principles for Research Software [2] released in 2022, initiatives for Software Management Plans (SMPs) [3, 4] and machine-actionable SMPs [5], and best practices [6, 7] or efforts to automatically mine software citations and create structured machine-interpretable knowledge about software usage [8]. The importance of research software is also recognized in the National Research Data Infrastructure (NFDI) in Germany with groups such as the NFDI-Research Software Engineer (NFDI-RSE).

2. Metadata for Research Software

Structured, semantic, and machine-actionable metadata is a must when it comes to the FAIR principles, either for data or software. Metadata makes it easier for aggregators, archives and registries to provide a quick and open overview even if the described object is not openly available. Metadata for research software should be simple and flexible, focusing on those common elements, across software produced in a variety of scientific disciplines. A good starting point are the FAIR principles as they already suggest some metadata elements such as identifier, license, provenance and meaningful links to related objects (e.g., data consumed and produced by a software); some additional elements can be taken from scholarly publications metadata (e.g., creators, contributors, keywords).

In recent years, different efforts have repurposed Schema.org [9] for its use in science. It is a vocabulary developed by a community involving major search engines, and offers a simple way for web pages to semantically describe their content by embedding structured markup. Bioschemas [10] and CodeMeta [11] build on top of Schema.org and provide specifications to describe research software. Bioschemas ComputationalTool specification is used in bio.tools [12], while CodeMeta is used in the Software Heritage Foundation Archive [13]. NFDI-RSE is currently working on a software common marketplace that would benefit from the use of community-agreed metadata as it would enable information retrieval for software across different disciplines. A common metadata layer for research software would also make it easier to interoperate with extended and richer versions used along the consortia. For instance, the NFDI4DataScience requires additional metadata to describe training and optimization processes done with software created and/or used in solutions using data science and artificial intelligence technologies.

3. Software Management Plans

Data Management Plans (DMPs) are text-based documents describing the data management lifecycle from collection to preservation. Machine-actionable DMPs (ma-DMPs) [14] add a structured layer on top, so it becomes easier to automate the integration of information and updates. Software Management Plans pursue the same aim but for software. For instance, the Software Best Practices Focus Group, part of the ELIXIR Tools Platform, proposed an SMP for Life Sciences [3]. Similarly, the Netherlands eScience Center and the Dutch Research Council (NWO) have developed (national) guidelines for domain-agnostic SMPs [4]. Same as DMPs, SMP templates commonly pose questions to ensure that researchers follow some minimum software management standards and policies when developing research software. SMP would also help in better understanding inner workings of software, thus providing ground for a better explanation of research outcomes.

To improve interoperability and reusability of SMPs, a machine-actionable version of the ELIXIR SMP is under development [5, 15]. This ma-SMP version builds on top of the ma-DMPs so they can be easily integrated with each other. It reuses and harmonizes elements from the ma-DMP, Schema.org, Bioschemas and CodeMeta specifications, while also adding new types and properties. An overview is shown in Figure 1. In terms of NFDI, machine-actionability for DMPs and SMPs would make it easier to connect them to each other, while also developing templates tailored to different communities with a common ground, making it easier to, for instance, compare plans across different consortia and disciplines.

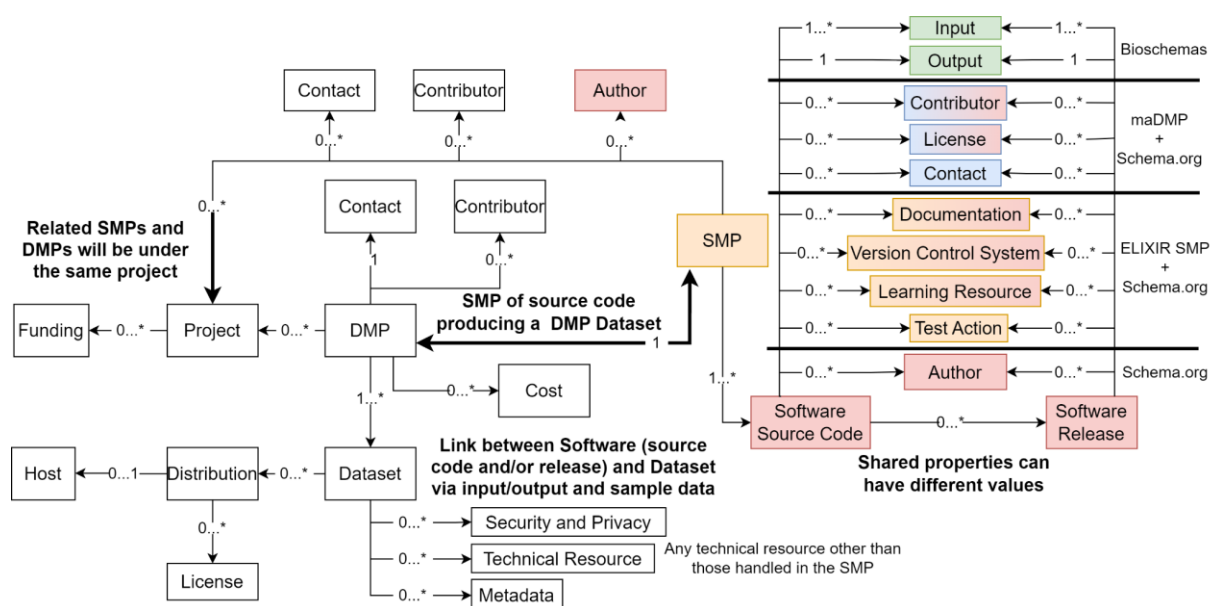


Figure 1. Metadata model for maSMP. Boxes with colored backgrounds correspond to the elements added for the maSMP case.

4. Future Work

Machine-actionability for DMPs should be embraced by the DMP working group part of NFDI-Infra section. Efforts should be combined with the RSE working group to also include maSMPs. In addition, NFDI consortia are working on the extraction of machine-interpretable metadata about software and their use in research, e.g. aiming at creating structured knowledge graphs of software and their scholarly adoption and use [8, 16]. While advancing the quality of information extraction baselines for such tasks is crucial to improve metadata quality, shared tasks on software mention detection and metadata extraction are currently being organized by the NFDI community. Using SchemaOrg as a lightweight gluing point seems reasonable as there are already efforts in that direction, it is domain-agnostic and can be customized following, for instance, the profile-way proposed by Bioschemas. By bringing multiple disciplines and communities together, NFDI is in a unique position to get community-based agreements wrt metadata for science.

Data availability statement

The metadata model corresponding to machine-actionable Software Management Plans has been published as an ontology and can be accessed at <https://doi.org/10.5281/zenodo.7806639>.

Author contributions

OG: conceptualization, project administration, writing – review & editing. DD: writing – review & editing. SD: writing – review & editing. DRS: conceptualization, writing – review & editing. LJC: conceptualization, funding acquisition, project administration, writing – original draft, writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

The machine-actional SMP project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017536 and is part of the Research Data Alliance and European Open Science Cloud Future call 2022. NFDI4DataScience consortium is funded by the Deutsche Forschungsgemeinschaft DFG, project no.460234259.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3: 160018. <https://doi.org/10.1038/sdata.2016.18>
2. Chue Hong, NP. et al. FAIR Principles for Research Software (FAIR4RS Principles). Research Data Alliance. 2022 <https://doi.org/10.15497/RDA00068>
3. Alves, R., Bampalíkis, D., Castro, L., Fernández, J. M., Harrow, J., Kuzak, M., ... Via, A. (2021, October 25). ELIXIR Software Management Plan for Life Sciences. <https://doi.org/10.37044/osf.io/k8znb>
4. Martínez-Ortiz C, Martínez Lavanchy P, Sesink L, Olivier BG, Meakin J, de Jong M, et al. Practical guide to Software Management Plans. Zenodo; 2022 Oct. <https://doi.org/10.5281/zenodo.7248877>
5. Giraldo O, Alves R, Bampalíkis D, Fernández González JM, Martín Del Pico E, Psomopoulos F, et al. A metadata analysis for machine-actionable Software Mng Plans - Poster. ZB MED - Informationszentrum Lebenswissenschaften; 2023. Available: <https://doi.org/10.4126/FRL01-006440396>
6. Scheliga KS, Pampel H, Konrad U, Fritzsche B, Schlauch T, Nolden M, et al. Dealing with research software: Recommendations for best practices. 2019. <https://doi.org/10.2312/os.helmholtz.003>
7. Jiménez RC, Kuzak M, Alhamdoosh M, Barker M, Batut B, Borg M, et al. Four simple recommendations to encourage best practices in research software. *F1000Res*. 2017;6: 876. <https://doi.org/10.12688/f1000research.11407.1>
8. Schindler D, Bensmann F, Dietze S, Krüger F. The role of software in science: a knowledge graph-based analysis of software mentions in PubMed Central. *PeerJ Computer Science* 8:e835. 2022. <https://doi.org/10.7717/peerj-cs.835>
9. Guha RV, Brickley D, Macbeth S (2016) Schema.org. *Communications of the ACM* 59 (2): 44- 51. <https://doi.org/10.1145/2844544>
10. Gray AJG, Goble C, Jimenez RC (2017) From Potato Salad to Protein Annotation. ISWC Posters and Demo session. URL: <http://ceur-ws.org/Vol-1963/paper579.pdf>
11. Boettiger C. et al. CodeMeta: Minimal metadata schemas for science software and code, in JSON-LD. (Version 2.0). Available at <https://github.com/codemeta/codemeta>
12. Ison, J. et al. (2015). Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkv1116>
13. Abramatic J-F, Di Cosmo R, Zacchiroli S. Building the universal archive of source code. *Commun ACM*. 2018;61: 29–31. <https://dl.acm.org/doi/10.1145/3183558>
14. Miksa T, Oblasser S, Rauber A. Automating Research Data Management Using Machine-Actionable Data Management Plans. *ACM Trans Manage Inf Syst*. 2021;13: 18:1-18:22. <https://doi.org/10.1145/3490396>
15. Giraldo O., Geist L., Quiñones N., Solanki D., Rebholz-Schuhmann D., and Castro LJ. Machine-actionable Software Management Plan Ontology (maSMP ontology) (Version 0.0.1) [Dataset]. <https://doi.org/10.5281/zenodo.7806639>
16. Schindler, D., Bensmann, F., Dietze, S., Krüger, F., SoMeSci—A 5 Star Open Data Gold Standard Knowledge Graph of Software Mentions in Scientific Articles, 30th ACM International Conference on Information & Knowledge Management (CIKM2021), ACM 2021. <https://doi.org/10.1145/3459637.3482017>

Determining the Similarity of Research Data by Using an Interoperable Metadata Extraction Method

Benedikt Heinrichs¹[\[https://orcid.org/0000-0003-3309-5985\]](https://orcid.org/0000-0003-3309-5985), and
M. Amin Yazdi¹[\[https://orcid.org/0000-0002-0628-4644\]](https://orcid.org/0000-0002-0628-4644)

¹IT Center, RWTH Aachen University, Seffenter Weg 23, Aachen, Germany

Abstract: Determining the similarity of research data is not a simple task, as the formats can differ widely depending on the domain. Especially, since many formats are represented as binary files, the raw comparison of these will not yield good results. This makes it hard to accurately tell how similar certain research work is by comparing the data. With the emergence of extracted interoperable metadata, a form to describe data has been provided which is independent of the data format. Therefore, this work tries to use this extracted interoperable metadata and create a method to determine the similarity of research data based on their metadata. The produced method utilizes domain knowledge about the extracted metadata and the way they are formulated. A baseline is created, and further methods are created to compare to. The results show that our method outperforms all other methods, especially the ones which are focused on comparing the research data itself, not the metadata. Since the results are promising, we propose further investigations against other datasets and possible use cases.

Keywords: Research Data Management, Data Similarity, Metadata Similarity, Linked Data

1 INTRODUCTION

Typically, researchers accumulate large datasets while conducting scientific studies, which are often stored in various ways and locations depending on the domain. With the emergence of research data management and the FAIR Principles, scholars demanding to make research data Findable, Accessible, Interoperable, and Reusable [1], the current trend is to deal with this divergence. However, even though research data management (RDM) can provide platforms where research data can be stored and put forward recommendations for certain file types based on the domain, the reality is that there is no one-for-all solution. This is mainly because the scientific disciplines differ wildly and, with that, their data as well (e.g. texts, measurements, Excel sheets, and code). Especially since researchers aim to deliver novel studies, the choice for file types occasionally ought to be of a unique type to satisfy the requirements of a study. However, this creates a challenge when it is necessary to compare research data and assess the pairwise relevancy of files. While with texts, this might be an achievable task [2], with more complex file types represented in binary formats, this comparison

gets complicated very fast [3]. With different file types, this comparison can become impossible if one only has the raw binary format. Thankfully, work has been conducted to extract the most descriptive parts of research data and describe it as interoperable metadata [4]. With this metadata, a standard way exists to summarize research data, and it can be represented as a graph. Since with this, a uniform description of research data can be created, the task here is to extend the previous work and see how well this extracted metadata can be used to determine the similarity of research data, regardless of their type. The goal in mind is that such a method should outperform methods that try to directly compare two binaries of research data. Additionally, the created method should be provided as open-source software so that it can be easily integrated into RDM workflows.

2 CONTRIBUTION

For the contribution, we will shortly go over the background, discuss our approach, and finally present our results.

2.1 Background

The used metadata extraction method [5] [4] makes use of software like Apache Tika [6] to create interoperable metadata. This interoperable metadata is described using the Resource Description Framework (RDF) [7] and makes use of ontologies like DCAT [8]. Since RDF metadata can be formulated as a graph, graph similarity methods are viable candidates for the similarity approach [9]. Especially, works about structural similarity [10] and entity comparison [11] provide some building blocks for the proposed research goal.

2.2 Approach

Our proposed similarity method makes use of interoperable metadata as a graph. Additionally, with our domain-knowledge, we identify several optimizations that we can apply to improve upon standard similarity metrics. These optimizations aim to *Filter* out unique subjects, utilize the metadata *Structure* given by DCAT, and make the output *Simpler* by removing specific relationship triples (FSS). These optimizations were applied to similarity metrics like Jaccard, Cosine, and a customized metric. Our favored metric for this domain is Jaccard, so with the optimizations, our proposed method was called FSS Jaccard. For a comparison, additional methods were created that do not follow these optimizations. Lastly, methods were built that compare the research data binaries itself and not the interoperable metadata. All of these methods were compared on an evaluation dataset containing research data. The implementation of these methods is provided in this open-source repository: <https://git.rwth-aachen.de/coscine/research/semanticssimilarity>. The code can be run as a service so that it is easily integrable in RDM workflows.

2.3 Results

For evaluating the proposed similarity method, it was put against the other comparison methods. They all ran on the same evaluation dataset, and the results were evaluated using specific test characteristics and reliability measures. They were chosen based on the works of [12], and [13]. The results can be found in figure 1.

	Error Values		Reliability	Test Characteristics			Inter-rater reliability
	Mean	SD	true score	Sensitivity	Specificity	Accuracy	Cohen's Kappa
FSS Similarity	0,021	0,0442	0,9633	0,6491	0,9655	0,9031	0,6677
FSS Cosine	0,0218	0,0445	0,9628	0,6491	0,9655	0,9031	0,6677
FSS Jaccard	0,0195	0,0436	0,9628	0,6721	0,9912	0,9239	0,7437
Filter Similarity	0,0495	0,0494	0,9404	0,8182	0,7607	0,7716	0,4386
Cosine Binary	0,1304	0,2061	0,6772	0,9459	0,5952	0,6401	0,2514
Jaccard Binary	0,0504	0,1253	0,8433	0,6596	0,8182	0,7924	0,3853
Jaccard Similarity	0,0264	0,0704	0,9451	0,5738	0,9912	0,9031	0,6601

Note. Error values are determined by the absolute distance of the values from the validation values. Reliability is defined as the proportion of true variance, measured as covariance between the validation values and the test values, of the overall variance of the test values. Values are considered true positive if the data and validation value is larger than 0, but the data value is not bigger than 1.3 times the validation value (false positive) or less than 77% of the validation value. The data is still considered true negative if the validation value is 0 and the data value is less than 0.083 (average standard deviation of all algorithms).

Figure 1. Diagnostic Analysis

What can be seen from figure 1 is that FSS Jaccard (our proposed method) performs better than all the other methods when looking at the error values, reliability and inter-rater reliability. Especially interesting is that all methods based on the interoperable metadata perform much better than the ones based on the binary itself (Jaccard Binary and Cosine Binary). We discuss that this is the case because the interoperable metadata has more useful data to compare to, while binaries are oftentimes very hard to compare. The results for our proposed method FSS Jaccard are visualized in figure 2.

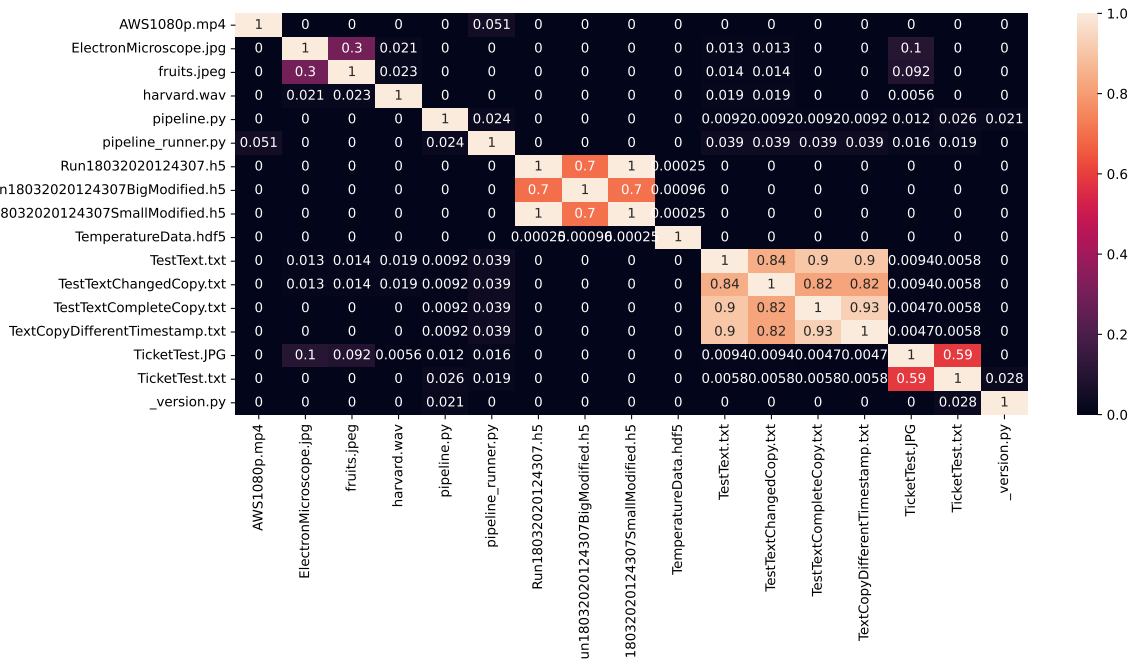


Figure 2. FSS Jaccard Results

The results in figure 2 show that there are not many outliers, and similar research data is usually identified as such. Especially interesting cases like smaller modifications between research data and research data which convey the same information but are in different file types (text and image) are accurately detected.

3 CONCLUSION

Our work established a new way to determine the similarity of research data based on interoperable metadata. First, we clarify our motivation, which ranges from the gen-

eral interest of figuring out how similar research data is to the problem of different data types. Then, we discuss the formal background and describe our similarity method. This method makes use of domain knowledge to craft a metric that filters the metadata, builds on top of the pre-existing structure, and simplifies it. The resulting algorithm is based on the Jaccard similarity metric, and we call it $FSS_{Jaccard}$. We determine other similarity metrics to which we want to compare our similarity metric to on an evaluation dataset. The results strongly suggest that $FSS_{Jaccard}$ outperforms the other similarity metrics on our limited evaluation dataset. This is due to the use of interoperable metadata and the utilization of domain knowledge. Especially, the similarity metrics based on the binary itself perform worse than the ones based on the interoperable metadata. With this, we conclude that it has some merit to compare research data using interoperable metadata. This comparison has additionally the benefit of being type-agnostic and can even achieve better similarity results, at least in our case.

Competing interests

The authors declare that they have no competing interests.

References

- [1] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data*, vol. 3, p. 160 018, 2016. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
- [2] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity—a survey," *ACM Comput. Surv.*, vol. 54, no. 2, Feb. 2021, ISSN: 0360-0300. DOI: [10.1145/3440755](https://doi.org/10.1145/3440755). [Online]. Available: <https://doi.org/10.1145/3440755>.
- [3] S. Kim, Y. J. Yoo, J. So, J. G. Lee, J. Kim, and Y. W. Ko, "Design and implementation of binary file similarity evaluation system," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 9, no. 1, pp. 1–10, 2014. DOI: [10.14257/ijmue.2014.9.1.01](https://doi.org/10.14257/ijmue.2014.9.1.01).
- [4] B. Heinrichs, N. Preuß, M. Politze, M. S. Müller, and P. F. Pelz, "Automatic General Metadata Extraction and Mapping in an HDF5 Use-case," in *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*, INSTICC, SciTePress, 2021, pp. 172–179, ISBN: 978-989-758-533-3. DOI: [10.5220/0010654100003064](https://doi.org/10.5220/0010654100003064).
- [5] B. Heinrichs and M. Politze, "Moving Towards a General Metadata Extraction Solution for Research Data with State-of-the-Art Methods," 12th International Conference on Knowledge Discovery and Information Retrieval, Nov. 2, 2020. DOI: [10.18154/RWTH-2020-12385](https://doi.org/10.18154/RWTH-2020-12385). [Online]. Available: <https://publications.rwth-aachen.de/record/809129>.
- [6] C. Mattmann and J. Zitting, *Tika in action*, 2011.
- [7] D. Wood, M. Lanthaler, and R. Cyganiak, "RDF 1.1 Concepts and Abstract Syntax," W3C, W3C Recommendation, Feb. 2014, <https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>.
- [8] A. Perego, A. G. Beltran, R. Albertoni, S. Cox, D. Browning, and P. Winstanley, "Data Catalog Vocabulary (DCAT) - Version 2," W3C, W3C Recommendation, Feb. 2020, <https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/>.
- [9] J. Carroll, "Matching rdf graphs," May 2002, pp. 5–15, ISBN: 978-3-540-43760-4. DOI: [10.1007/3-540-48005-6_3](https://doi.org/10.1007/3-540-48005-6_3).

- [10] P. Maillot and C. Bobed, "Measuring structural similarity between rdf graphs," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, ser. SAC '18, Pau, France: Association for Computing Machinery, 2018, pp. 1960–1967, ISBN: 9781450351911. DOI: [10.1145/3167132.3167342](https://doi.org/10.1145/3167132.3167342). [Online]. Available: <https://doi.org/10.1145/3167132.3167342>.
- [11] A. Petrova, E. Sherkhonov, B. Cuenca Grau, and I. Horrocks, "Entity comparison in rdf graphs," in *The Semantic Web – ISWC 2017*, C. d'Amato, M. Fernandez, V. Tamma, et al., Eds., Cham: Springer International Publishing, 2017, pp. 526–541, ISBN: 978-3-319-68288-4. DOI: [10.1007/978-3-319-68288-4_31](https://doi.org/10.1007/978-3-319-68288-4_31).
- [12] M. Eid, M. Gollwitzer, and M. Schmitt, *Statistik und Forschungsmethoden, Lehrbuch (Grundlagen Psychologie)*, ger, 3., korrigierte Auflage, Online-Ausgabe. Weinheim ; Basel: Beltz, 2013, 1 Online–Ressource (XXXII, 1024 Seiten), ISBN: 978-3-621-27524-8. [Online]. Available: https://content-select.com/index.php?id=bib_view&ean=9783621278348.
- [13] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and Psychological Measurement*, vol. 33, pp. 613–619, 1973. DOI: [10.1177/00131644730330030](https://doi.org/10.1177/00131644730330030).

Which FAIR Are You?

A Detailed Comparison of Existing FAIR Metrics in the Context of Research Data Management

Mario Moser¹[\[https://orcid.org/0000-0001-9325-4074\]](https://orcid.org/0000-0001-9325-4074),
Jonas Werheid¹[\[https://orcid.org/0009-0003-6022-2633\]](https://orcid.org/0009-0003-6022-2633),
Tobias Hamann¹[\[https://orcid.org/0000-0002-8021-5524\]](https://orcid.org/0000-0002-8021-5524),
Anas Abdelrazeq¹[\[https://orcid.org/0000-0002-8450-2889\]](https://orcid.org/0000-0002-8450-2889), and
Robert H. Schmitt^{1 2}[\[https://orcid.org/0000-0002-0011-5962\]](https://orcid.org/0000-0002-0011-5962)

¹Laboratory for Machine Tools and Production Engineering (WZL) of RWTH Aachen University, Germany

²Fraunhofer Institute for Production Technology (IPT), Aachen, Germany

Abstract: In data management the high-level FAIR principles are interpreted and implemented in various FAIR metrics. While this specific interpretation is intended, it leads to the situation of several metrics with different evaluation results for the same digital object. This work conducts an organizational-formal comparison, showing up elements like categories of importance in the considered metrics, as well as a content-wise comparison of selected metrics how they differ in their interpretation. The results give orientation especially to everyone in science aiming to find the right metric to make their data FAIR.

Keywords: Research Data Management - RDM - FDM; FAIR principles; FAIR metrics - comparison

1 Introduction

The FAIR principles [1] have been widely adopted as a guideline for making scientific data and scholarly digital objects more findable, accessible, interoperable, and reusable. Scholarly digital objects include data in a narrow sense as well as software, repositories, and workflows among others [1], [2]. The four foundational principles are described in more detail in fifteen guiding principles. These principles are interpreted and implemented in several FAIR metrics in order to meet domain-specific requirements or to focus on special types of digital objects [3]. This poses challenges for several addressees in research: scientists might want to select a metric to design their digital objects FAIR; initiatives working on developing new (specific) metrics need to understand existing ones and identify gaps; and research funding agencies might want to select an appropriate metric to measure the FAIRness of published results. Therefore, a detailed comparison of FAIR metrics is necessary to investigate differences and how they affect the data management and evaluation results for digital objects.

Various FAIR metrics are used in FAIR assessment tools [4]. While comparisons of FAIR assessment tools and their outcomes already exist (e.g. [5], [6], [7], [8], [9], [10]), they usually do not focus on the underlying metrics themselves as the reasons for differences observed. At least the tool evaluation in [11] includes comparison of the metrics Maturity Indicators (MI) and FAIRsFAIR (FsF). The tool comparison in [12] mentions the used metrics as one aspect. The development of the European Open Science Cloud (EOSC) FAIR metrics [13] made comparisons to metrics developed previously, but no complete overview comparing several metrics is currently available.

In our comparison, metrics that satisfies the following conditions are selected: Metrics that

- are generally applicable, i.e. that are not limited to one specific type of digital objects, since this reduces comparability;
- follow the structure of the fifteen FAIR guiding principles; and
- have a separate documentation following the characteristics in [14]’s table 1, such as a metric identifier and name as well as what and how is it measured.

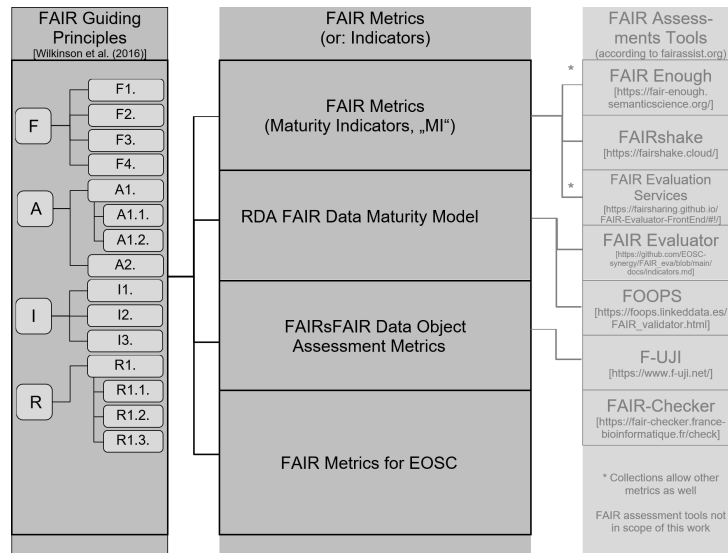
Metrics are identified based on scientific literature research as well as on considering metrics used in publicly available FAIR assessment tools. (Although the assessment tools themselves are mentioned in this paper, they are not compared.) For the identified metrics the comparison divides into two types: 1) An organizational-formal comparison focuses on characteristics like the evaluated digital objects, evaluation scale and metric item weighting factors. 2) Starting from each guiding principle a one-by-one content-wise comparison of each FAIR metric is conducted among the other metrics as well as with the (general) guiding principles itself.

2 Results

Four general FAIR metrics are identified for this work, presented in the middle column in figure 1, including their optional usage in assessment tools. As stated before, specific metrics are excluded (e.g. for FAIRness of software: [15], FAIR4RS ([16], [17]), [18]). The FAIRplus (F+) indicators [19] as well as several self-assessment checklists have been not been incorporated due to a missing structural similarity to the FAIR guiding principles.

The organizational-formal comparison reveals that metrics focus on digital objects in general (MI, RDA, FsF) or more specific on the web (EOSC). This outcome is not surprising since the concept of FAIR metrics is to specify the general FAIR principles for a certain usage. While e.g. the FAIR MIs are generally applicable, there is another metric rephrased to focus explicitly on software. The assessment tools “FOOPS” uses the (general) RDA metric for the evaluation of ontologies. RDA prioritizes the relevance of their metric items in three categories (“useful”, “important” and “essential”). Following the design framework template by [14] reveals room for improvement in documentation.

The content-wise comparison of FAIR metric elements shows different numbers of metric items per metric: While the MIs are with 14 items quite close to the FAIR guiding principles, on the opposite site the RDA contains 41 items in their metric. The different number of metric items is mainly caused by splitting into multiple and more fine-granular metric items for one guiding principle: Data vs. metadata are distinguished; principles are divided into multiple parts (e.g. RDA-I1-01: “uses knowledge representation expressed in standardised format” and in addition “uses machine-understandable knowledge representation”); principles (e.g. A1.) with sub-principles (e.g. A1.1., A1.2.)



MI [20] [21], RDA [22] [23], FAIRsFAIR [24], EOSC [13]

Figure 1. FAIR metrics identified by usage in FAIR assessment tools

are either described on all levels or on subprinciple level only; human as well as machine readability is in some metrics explicitly distinguished. The EOSC that has marked some items “temporarily”, which might evolve over time. The development of the EOSC FAIR metric has been influenced by the RDA’s and FAIRsFAIR’s metrics, what is expressed in similarity of the content. E. g. RDA-F1-01D demands that “Data is identified by a persistent identifier” [22] while FsF-F1-02D in a similar way states that “Data is assigned a persistent identifier” [24]. Therefore the RDA FAIR Data Maturity Model has an extended (direct/indirect) influence on the development of the presented subsequent metrics by FAIRsFAIR and EOSC. A high-level overview of the above-mentioned results is shown in figure 2.

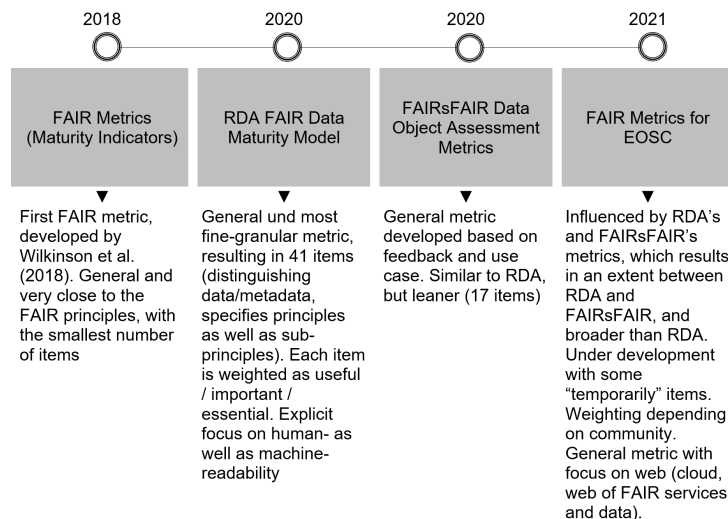


Figure 2. Simplified FAIR metrics overview with the results from the comparison

3 Outlook

Our detailed comparison of FAIR metrics reveals the differences on an organizational-formal level as well as content-wise. Differences are thematised and thus show the (intended) different interpretation and implementations of the FAIR principles.

The outcome can be one part of the explanation why different FAIR assessment tools lead to different results for the same digital object. For scientists, initiatives working on FAIR, and research funders, the results might be useful to decide which FAIR metric to use in their work. When developing a new metric, the comparison can help to identify gaps or to check typical design elements of metrics (like weighting categories, aspects to (not) focus on etc.).

It has to be mentioned that this analysis represents the current situation at the moment of execution. Due to the continuous evolution of FAIR, an ongoing evaluation of differences will be required. Existing metrics might change, and new metrics might be developed for specific purposes. This is necessary to adapt to new future developments and requirements. The selection of metrics in this work has been rather strict, so more metrics exist that are not covered here. But more important than *which* FAIR you use is *that* your data is somehow FAIR.

Author contributions

Mario Moser: Conceptualization, Investigation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing;

Jonas Werheid: Investigation, Formal analysis, Writing – original draft;

Tobias Hamann: Supervision, Writing – review & editing;

Anas Abdelrazeq: Supervision, Writing – review & editing;

Robert Schmitt: Funding acquisition, Supervision

Competing interests

The authors declare that they have no competing interests.

Funding

The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.

Acknowledgements

The authors would like to thank the reviewers for their valuable and constructive feedback.

References

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, no. 160018, Mar. 2016. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).

- [2] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan, "Research Objects: Towards Exchange and Reuse of Digital Knowledge," *Nature Precedings*, Jul. 2010. DOI: [10.1038/npre.2010.4626.1](https://doi.org/10.1038/npre.2010.4626.1).
- [3] A. Jacobsen, R. de Miranda Azevedo, N. Juty, et al., "FAIR Principles: Interpretations and Implementation Considerations," *Data Intelligence*, vol. 2, no. 1-2, pp. 10–29, Jan. 2020. DOI: [10.1162/dint_r_00024](https://doi.org/10.1162/dint_r_00024).
- [4] A. Devaraju and R. Huber, "An automated solution for measuring the progress toward FAIR research data," *Patterns*, vol. 2, no. 11, Nov. 2021. DOI: [10.1016/j.patter.2021.100370](https://doi.org/10.1016/j.patter.2021.100370).
- [5] M. D. Wilkinson, S.-A. Sansone, G. Marjan, J. Nordling, R. Dennis, and D. Hecker, "FAIR Assessment Tools: Towards an "Apples to Apples" Comparisons," *Zenodo*, Dec. 2022. DOI: [10.5281/zenodo.7463421](https://doi.org/10.5281/zenodo.7463421).
- [6] K. Peters-von Gehlen, H. Höck, A. Fast, D. Heydebreck, A. Lammert, and H. Thiemann, "Recommendations for Discipline-Specific FAIRness Evaluation Derived from Applying an Ensemble of Evaluation Tools," *Data Science Journal*, vol. 21, no. 7, Mar. 2022. DOI: [10.5334/dsj-2022-007](https://doi.org/10.5334/dsj-2022-007).
- [7] N. Krans, A. Ammar, P. Nymark, E. Willighagen, M. Bakker, and J. Quik, "FAIR assessment tools: evaluating use and performance," *NanoImpact*, vol. 37, p. 100 402, 2033. DOI: [10.1016/j.impact.2022.100402](https://doi.org/10.1016/j.impact.2022.100402).
- [8] C. Bahim, M. Dekkers, and B. Wyns, "Results of an Analysis of Existing FAIR Assessment Tools," *Zenodo*, May 2019. DOI: [10.15497/rda00035](https://doi.org/10.15497/rda00035).
- [9] E. González, A. Benítez, and D. Garijo, "FAIROs: Towards FAIR Assessment in Research Objects," in *Linking Theory and Practice of Digital Libraries*, G. Silvello, O. Corcho, P. Manghi, et al., Eds., Cham: Springer International Publishing, 2022, pp. 68–80. DOI: [10.1007/978-3-031-16802-4_6](https://doi.org/10.1007/978-3-031-16802-4_6).
- [10] D. Slamkov, V. Stojanov, B. Koteska, and A. Mishev, "A Comparison of Data FAIRness Evaluation Tools," in *Ninth Workshop on Software Quality Analysis, Monitoring, Improvement, and Applications (SQAMIA 2022)*, Oct. 2022. [Online]. Available: <https://www.researchgate.net/publication/364308377>.
- [11] C. Sun, V. Emonet, and M. Dumontier, "A comprehensive comparison of automated FAIRness Evaluation Tools," *Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4HCLS)*, Dec. 2022, Additional material at <https://doi.org/10.5281/zenodo.5539823>. [Online]. Available: <https://ceur-ws.org/Vol-3127/paper-6.pdf>.
- [12] E. Kontsioti. "The Road to FAIRness: An Evaluation of FAIR Data Assessment Tools." (2023), [Online]. Available: <https://www.thehyve.nl/articles/evaluation-fair-data-assessment-tools> (visited on 04/26/2023).
- [13] European Commission and Directorate-General for Research and Innovation, J. M. Aronsen, O. Beyan, et al., *Recommendations on FAIR metrics for EOSC*, S. Jones and F. Genova, Eds. Publications Office, 2021. DOI: [10.2777/70791](https://doi.org/10.2777/70791). [Online]. Available: <https://op.europa.eu/o/opportal-service/download-handler?identifier=ced147c9-53c0-11eb-b59f-01aa75ed71a1&format=pdf&language=en&productionSystem=cellar&part=>.
- [14] M. D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L. O. Bonino da Silva Santos, and M. Dumontier, "A design framework and exemplar metrics for FAIRness," *Scientific Data*, vol. 5, no. 1, Jun. 2018. DOI: [10.1038/sdata.2018.118](https://doi.org/10.1038/sdata.2018.118).
- [15] A.-L. Lamprecht, L. Garcia, M. Kuzak, et al., "Towards FAIR principles for research software," *Data Science*, vol. 3, no. 1, pp. 37–59, Jun. 2020. DOI: [10.3233/DS-190026](https://doi.org/10.3233/DS-190026).

- [16] N. P. Chue Hong, D. S. Katz, M. Barker, *et al.*, "FAIR Principles for Research Software (FAIR4RS Principles)," *Zenodo*, May 2022. DOI: [10.15497/RDA00068](https://doi.org/10.15497/RDA00068).
- [17] M. Barker, N. P. Chue Hong, D. S. Katz, *et al.*, "Introducing the FAIR Principles for research software," *Scientific Data*, vol. 9, no. 622, pp–pp, Oct. 2022. DOI: [10.1038/s41597-022-01710-x](https://doi.org/10.1038/s41597-022-01710-x).
- [18] D. S. Katz, M. Gruenpeter, and T. Honeyman, "Taking a fresh look at FAIR for research software," *Patterns*, vol. 2, no. 1, Mar. 2021. DOI: [10.1016/j.patter.2021.100222](https://doi.org/10.1016/j.patter.2021.100222).
- [19] FAIRplus. "FAIRplus Indicators V0.1." (Oct. 2020), [Online]. Available: <https://fairplus.github.io/fairification-results/2020-10-11-FAIRplus-indicators-v0.1/> (visited on 04/26/2023).
- [20] M. D. Wilkinson, M. Dumontier, S.-A. Sansone, *et al.*, "Evaluating FAIR maturity through a scalable, automated, community-governed framework," *Scientific Data*, vol. 6, no. 174, Sep. 2019. DOI: <https://doi.org/10.1038/s41597-019-0184-5>.
- [21] FAIR Metrics Group. "GitHub FAIRMetrics." (2022), [Online]. Available: <https://github.com/FAIRMetrics/Metrics> (visited on 04/26/2023).
- [22] RDA FAIR Data Maturity Model Working Group, "FAIR Data Maturity Model: specification and guidelines," 2020. DOI: [10.15497/rda00050](https://doi.org/10.15497/rda00050).
- [23] C. Bahim, C. Casorrán-Amilburu, M. Dekkers, *et al.*, "The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments," *Data Science Journal*, vol. 19, no. 1, pp. 1–7, Oct. 2020. DOI: [10.5334/dsj-2020-041](https://doi.org/10.5334/dsj-2020-041).
- [24] A. Devaraju, R. Huber, M. Mokrane, *et al.*, "FAIRsFAIR Data Object Assessment Metrics (0.5)," *Zenodo*, Apr. 2022. DOI: [10.5281/zenodo.6461229](https://doi.org/10.5281/zenodo.6461229).

The RDM System LARA

- semantics through automation from bottom up

Mark Doerr¹[\[https://orcid.org/0000-0003-3270-6895\]](https://orcid.org/0000-0003-3270-6895), Stefan T. Maak¹[\[https://orcid.org/0009-0000-9930-0182\]](https://orcid.org/0009-0000-9930-0182),
Marian J. Menke¹[\[https://orcid.org/0000-0002-5887-3045\]](https://orcid.org/0000-0002-5887-3045), and
Uwe T. Bornscheuer¹[\[https://orcid.org/0000-0003-0685-2696\]](https://orcid.org/0000-0003-0685-2696)

¹University Greifswald, Greifswald, Germany

Abstract: LARAsuite (<https://gitlab.com/larasuite>) is a free and open source research data management system that addresses the problematic of manual data insertion and metadata assignment into the corresponding databases by a radically automated processes. Data and Metadata is mainly not inserted by humans, but by the machines producing the data and automatically transferring this generated data and metadata to the LARA RDM system. This data transfer is achieved through the intensive utilisation of the free and open lab-automation standard SiLA (Standardisation in Labautomation) (<https://sila-standard.org>), combined with a simple to note process description language (pythonLab) and orchestration, scheduling, data aggregation and evaluation. Many LARA instances can be selectively synchronised to form a decentralised network of data infrastructures across labs and institutions. Research data and meta-data can be queried by a SPARQL endpoint. As the LARAsuite comes close to an ideal FAIR-principles based RDM System¹, is generic to the most common natural science applications, is open source (python based), modular and easy deployable, it can be used to showcase NFDI goals for a wide range of scientists.

Keywords: RDMS, Semantics, Ontologies, Automation, SiLA, JSON-LD, Workflow-description

1 Introduction

1.1 Motivation of the LARA system

The majority of recent research data management systems (RDMS) are suffering from the fact that most data and metadata - and it's meaning (semantics) need to be entered manually. Also all relations between the data needs to be assigned by the hand of the researcher. The LARAsuite[2] addresses this by a radical automation approach. It uses the free and open source lab communication standard SiLA[3] and the underlying wire protocol gRPC.

¹LARAsuite tries to achieve full maturity levels according to the *RDA FAIR data maturity model Working Group's* maturity levels F1-4, A1-2, I1-3 and R1 (1)

1.2 Key features of the LARA system

- fully open source and (mostly) written in python - so any scientist can adjust the system to the needs
- modular and extensible due to its modular architecture / (partly) micro-services new functionality can be added
- easy deployable through docker containers / docker-compose or kubernetes

1.3 Workflow with the LARA system

Projects and experiments are designed with a web-interface or a python notation. Processes can be formulated in the pythonLab process description language [4], which is interpreted by the LabOrchestrator [5]. The LabOrchestrator orchestrates the whole experiment, executed by machines and humans (also in collaboration between humans and machines). Which parts of processes are executed by which entity (machine or human) is described in the *pythonLab procedures* defining how a process is executed. Humans have always the option to intervene any running process by pausing, interrupting / resuming it². For the scheduling of all the experimental steps the LabScheduler[6] can be used. All data is collected in the LARA databases and automatically semantically enriched, using modular ontologies. Data can then automatically evaluated and visualised or further processed by interfacing to the jupyter infrastructure.

2 Architecture of LARA

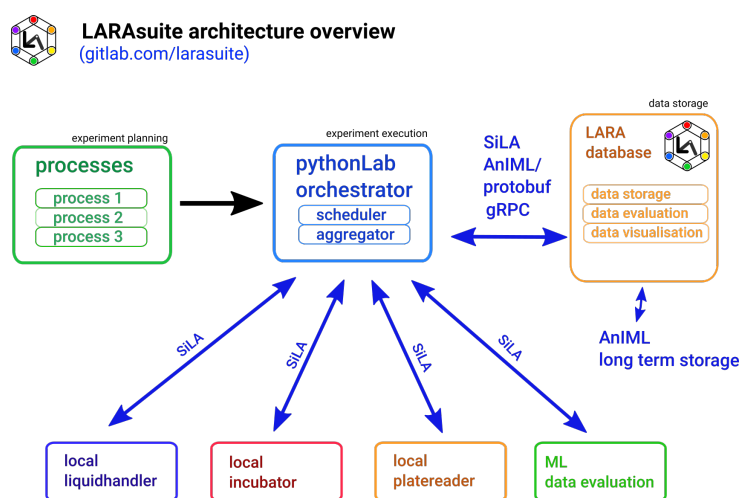


Figure 1. The LARA architecture.

The LARA suite has currently components to model the following entities of projects and experiments

- Projects and Experiments
- User
- Processes
- Data (including associated Metadata)
- Material (Parts, Devices, Device Setups, Labware)- including an inventory
- Substances (all chemical substances and mixtures) - including an inventory

²Modification of a running process is currently under development

- Sequences (in case substances have a sequence, like DNA, RNA, plasmids, peptides, proteins,
- Organism (Bacteria, Plants, Animals, ...)

More details can be found on the LARAsuite Webpage: <https://github.com/larasuite/lara>.

Data availability statement

The full source code of the LARAsuite [2] is available on gitlab: <https://gitlab.com/larasuite> - GPL-3 licensed. The ontologies, semantic web tools and SiLA servers are also freely available <https://gitlab.com/opensourclab>[7].

Underlying and related material

Videos of the University LARA robotic platform, which runs with the LARAsuite, can be viewed here: <https://lara.uni-greifswald.de>

Author contributions

Conceptualization: M.D.; Methodology: M.D, S.T.M; Software: M.D, S.T.M; Writing – Original Draft: M.D., Writing – Review Editing: M.D.; Visualization: M.D.; Project-Funding: U.B., M.D.

Competing interests

The authors declare that they have no competing interests.

Funding

The Deutsche Forschungsgemeinschaft (DFG) is acknowledged for funding this research as part of the Nationale Forschungsdateninfrastruktur (NFDI) initiative (grant No.: NFDI/2-1 - 2021).

The Bundesministerium für Forschung und Wissenschaft (BMBF) - KIWI-biolab project.

Acknowledgements

We would like to acknowledge Stefan Born (TU-Berlin) for his very valuable comments.

References

- [1] R. F. data maturity model Working Group. "Fair data maturity model: Specification and guidelines." (2023), [Online]. Available: <https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines-0> (visited on 07/25/2023).
- [2] M. Doerr. "The lara suite." (2023), [Online]. Available: <https://github.com/LARAsuite> (visited on 04/26/2023).
- [3] T. S. organisation. "Sila - standards in lab automation." (2023), [Online]. Available: <https://sila-standard.org> (visited on 04/26/2023).

- [4] M. Doerr. "The pythonlab process description language." (2023), [Online]. Available: <https://github.com/opensourcelab/pythonlab> (visited on 04/26/2023).
- [5] M. Doerr and S. T. Maak. "The laborchestrator." (2023), [Online]. Available: <https://github.com/opensourcelab/laborchestrator> (visited on 04/26/2023).
- [6] S. T. Maak and M. Doerr. "The labscheduler." (2023), [Online]. Available: <https://github.com/opensourcelab/labscheduler> (visited on 04/26/2023).
- [7] M. Doerr. "The opensource lab project." (2023), [Online]. Available: <https://github.com/OpenSourceLab> (visited on 04/26/2023).

Open Science and Language Data: Expectations vs. Reality

The Role of Research Data Infrastructures

Paweł Kamocki¹[\[https://orcid.org/0000-0003-4881-7549\]](https://orcid.org/0000-0003-4881-7549), Erhard Hinrichs²[\[https://orcid.org/0009-0006-4192-7779\]](https://orcid.org/0009-0006-4192-7779), Sabine Springer³[\[https://orcid.org/0009-0004-8395-4279\]](https://orcid.org/0009-0004-8395-4279), Peter Leinen⁴[\[https://orcid.org/0000-0002-3014-000X\]](https://orcid.org/0000-0002-3014-000X), Andreas Witt⁵[\[https://orcid.org/0000-0002-0299-5713\]](https://orcid.org/0000-0002-0299-5713), and Dorothea Zechmann⁶[\[https://orcid.org/0009-0008-2704-2135\]](https://orcid.org/0009-0008-2704-2135)

^{1; 2; 5} Leibniz-Institut für Deutsche Sprache, Mannheim, Germany

^{3; 4; 6} Deutsche Nationalbibliothek, Germany

Abstract. Language data are essential for any scientific endeavor. However, unlike numerical data, language data are often protected by copyright, as they easily meet the threshold of originality. The role of research infrastructures (such CLARIN, DARIAH, and Text+) is to bridge the gap between uses allowed by statutory exceptions and the requirements of Open Science. This is achieved on the one hand by sharing language data produced by research organisations with the widest possible circle of persons, and on the other by mutualizing efforts towards copyright clearance and appropriate licensing of datasets.

Keywords: Research infrastructures, Text data, Open science

Language data are essential for any scientific endeavor. Since natural language is a primary tool for human communication, the importance of language data reaches beyond language science and the humanities, also to the STEM. The importance of language data has been highlighted by the recent boom in generative chatbots (such as Chat GPT), based on Large Language Models (LLMs). However, unlike numerical data, language data are often protected by copyright, as they easily meet the threshold of originality (in 2009, the Court of Justice of the European Union ruled that texts as short as 11 words can be protected by copyright [1]). This means that, in principle, language data cannot be reproduced (copied) and communicated to the public (shared) without permission from the rightholders, unless in cases expressly authorised by statutory provisions (known as exceptions or limitations). The development of LLMs by large and mostly US-based companies was allowed under legal frameworks that do not apply in Europe (such as the US fair use doctrine).

At the beginning of copyright history, and for a relatively long time, research activities were regarded as irrelevant from the point of view of copyright, or exempted under the ‘de minimis’ principle. It is only in the second half of the 20th century, with the development of reproduction technologies (such as Xerox machines) that the conflict between the interests of copyright holders and the needs of academia became apparent, and research exceptions were introduced in national, and in 2001 also in European copyright law. For years, research exceptions could not keep up with technological transformation of science, and they were considered outdated and insufficient. Things gradually changed first at the national level, with the adoption of the German *Gesetz zum Urheberrecht für die Wissenschaft* (UrhWissG, entered into force in 2018)[2], and then also at the European level, with the 2019 Directive on Copyright in the Digital Single Market (transposed in Germany in 2021)[3].

The German text contained a relatively robust exception for Text and Data Mining (TDM) for non-commercial research purposes; it also allowed the German National Library to build so-called citation archives of publicly available works for research purposes (§16a DNBG).

The European text then harmonized exceptions for TDM for scientific research purposes. Under current German law (§60d UrhG), research organisations can make reproductions of any material that they have lawful access to for TDM purposes. Moreover, the resulting corpus can be shared with a limited circle of persons for joint scientific research (which, rather regrettably, does not seem to be the case in all EU Member States).

Although the TDM exception was a welcome development, it accentuated the gap between what is allowed by copyright law, and what is required by the principles of Open Science, which today is of fundamental importance for any well-managed and ethical research project. Open Science requires open availability of research data, i.e. the possibility for everyone to reuse the data for any purpose[4]. *Vis-à-vis* this requirement, the possibility to share research data with 'a limited circle of persons for joint scientific research' is clearly insufficient. In that sense, the TDM exception failed to meet the needs of the research community; it may even have an adverse effect of creating isolated data ponds rather than robust knowledge commons by disincentivizing proper licensing of research data. Such proper licensing requires time and effort (which does not always produce the desired result), and it may be tempting for researchers to just rely on the statutory exception and settle on making their data available only to a limited circle of project partners instead.

The problem has certainly been noticed by both the German and the European legislators, as attested by proposed legislation such as the Research Data Act (*Forschungsdatengesetz*) or, at the European level, the Data Act, both intended to grant researchers access to data held by the private sector. The creation of Common European Data Spaces (including the Common European Language Data Step) is already a step in this direction. In this new paradigm the weight is shifted from property interests in data (such as copyright, database right, trade secret) to data governance and public interest (rights of access, portability), with the intention to rebalance the data market. One can point out that these efforts are not necessarily a step towards real Open Science, as the beneficiaries of the access right will certainly be limited (e.g., to public research institutions).

While these proposed developments are welcome they are still quite far from becoming a reality. For now, research infrastructures such as CLARIN, DARIAH, and Text+ are essential for providing researchers (and not only) with access to text data.

Text+ is a consortium of the national research data infrastructure (*Nationale Forschungsdateninfrastruktur*, NFDI). Text+ is not limited to existing data, but will systematically expand its portfolio in close consultation with the expert communities involved. This also includes tools to support researchers in the FAIR creation, use and provision of data throughout the entire data lifecycle.

Although the amount of data held by language data infrastructures is small compared to data used to develop the latest LLMs, their comparative advantage lies in the quality and heterogeneity. Data used to train LLMs are mostly obtained via web scraping, so their quality is low compared to carefully selected, annotated and curated multilingual language resources held by European infrastructures. One can hypothesise that this superior quality could counterbalance quantitative limitations, and allow to build leaner, less energy- and compute-hungry, but equally performant language models.

From the perspective of intellectual property rights, the role of research infrastructures is to bridge the gap between uses allowed by statutory exceptions and the requirements of Open Science. This is achieved on the one hand by sharing language data produced by research organisations with the widest possible (yet still 'strictly limited') circle of persons, and

on the other by mutualizing efforts towards copyright clearance and appropriate licensing of datasets.

Competing interests

The authors declare that they have no competing interests.

References

1. Court of Justice of the European Union, Case C-5/08, Infopaq International A/S v. Danske Dagblades Forening, 16 July 2009.
2. Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft vom 1. September 2017, Bundesgesetzblatt Jahrgang 2017 Teil I Nr. 61, pp. 3346-3351.
3. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, PE/51/2019/REV/1, OJ L 130, 17.5.2019, p. 92–125.
4. UNESCO Recommendation on Open Science, UNESCO 2021.

Object-related Research Data Workflows within NFDI4Objects and beyond

Florian Thiery¹[\[https://orcid.org/0000-0002-3246-3531\]](https://orcid.org/0000-0002-3246-3531),
Allard W. Mees¹[\[https://orcid.org/0000-0002-7634-5342\]](https://orcid.org/0000-0002-7634-5342),
Bernhard Weisser²[\[https://orcid.org/0000-0001-5262-2731\]](https://orcid.org/0000-0001-5262-2731),
Felix F. Schäfer²[\[https://orcid.org/0000-0002-9867-5588\]](https://orcid.org/0000-0002-9867-5588),
Stefanie Baars²[\[https://orcid.org/0000-0003-1100-6494\]](https://orcid.org/0000-0003-1100-6494),
Sonja Nolte³[\[https://orcid.org/0009-0002-1843-8791\]](https://orcid.org/0009-0002-1843-8791),
Henriette Senst⁴[\[https://orcid.org/0000-0003-2255-7478\]](https://orcid.org/0000-0003-2255-7478), and
Philipp von Rummel⁴[\[https://orcid.org/0000-0001-7545-2181\]](https://orcid.org/0000-0001-7545-2181)

¹Leibniz-Zentrum für Archäologie, Mainz, Germany

²Stiftung Preußischer Kulturbesitz, Berlin, Germany

³Verbundzentrale des GBV, Göttingen, Germany

⁴Deutsches Archäologisches Institut, Berlin, Germany

Abstract: n.a.

Keywords: NFDI, NFDI4Objects, Object Biography, Digitalisation Workflow, Semantic Modelling, Linked Open Data, Wikidata

1 Introduction

NFDI4Objects (N4O) represents a broad community dealing with material remains of human history from around 3 million years and involves numerous disciplines from the humanities, cultural studies and natural sciences with an archaeological and historical focus [1]. The objects examined include potsherds of common ware, artworks such as sculptures or jewellery, serially produced objects such as coins, organic remains such as wood, bones or pollen, inscribed clay tablets, papyri and stones, architectural remains, as well as human-modified landscapes. Modern research materials such as plaster casts, analogue photographs and drawings, archival documents, books and raw digital data are equally relevant.

2 Object biographies as challenges for RDM

Objects are in multiple relationships in terms of both their materiality and the actors, spaces and times associated with them. Within these networks, they are not static but are subject to change, which imbue them with individual biographies that can only partially be understood and reconstructed. Objects can appear in different contexts: the production by specific actors, the primary or secondary use in, for example, religious, military or sepulchral settings, their place of discovery as the physical context

set against their reception in the modern collection, mediation and research environments. Objects are constantly changing through use, decay and restoration. In the course of an excavation, their context is often irreversibly destroyed. In collections, the objects are often recombined and rearranged and reinterpreted according to changing research paradigms. The information about their find circumstances, preservation conditions and other contexts are decisive for their interpretation.

The NFDI4Objects' four-fold central task includes: a) comprehending the representations of these physical, three-dimensional objects as research data, b) relating them to the respective fluctuating contexts, c) transforming them adequately into digital space, and d) curating them according to subject-specific requirements. This involves overcoming various media breaks. Language, texts and images offer opportunities to bridge the media transfer. Machine-readable interfaces using community standards also enable computer-aided use. However, many uncertainties remain, often not sufficiently considered in the digital space. The digital FDM of objects opens up the potential for digital exploration of an object's diverse stories and meanings and for establishing further links.

This paper demonstrates the scope and vision of N4O and the challenges of an object-related RDM, exemplified by an object group of ancient pottery (African Red Slip Ware). It also addresses the questions and structures of further development and the transferability to other NFDI domains.

3 The African Red Slip Ware as Paradigm

An example of such archaeological objects is the African Red Slip Ware pottery (ARS), a category of fine ware produced in the Roman Imperial period [2], [3] (Figure 1). A feature of ARS is its applied relief decoration, which displays, e.g., mythological, Old and New Testament motifs, circus, arena and hunting scenes [4]. Thus, ARS is one indicator of exchange processes in the Roman empire.



Figure 1. (Left) O.39675 "bowl with man and woman"; (right) Various ARS objects of the LEIZA collection. [left: object LEIZA, photo: Lübke & Wiedemann, Leonberg; left: CC BY-SA 4.0 ARS3D Project / i3mainz / LEIZA, published in [5]]

All German ARS collections comprise numerous objects originating from private collections or were acquired on the art market before the KultGüRückG (Act on the Return of Cultural Property) was implemented in 2007 [6]–[8]. Thus the acquisition history inevitably tackles ethical issues, thus being related to the CARE principles. This requires a historical investigation of the provenance and a critical debate about this research data [9].

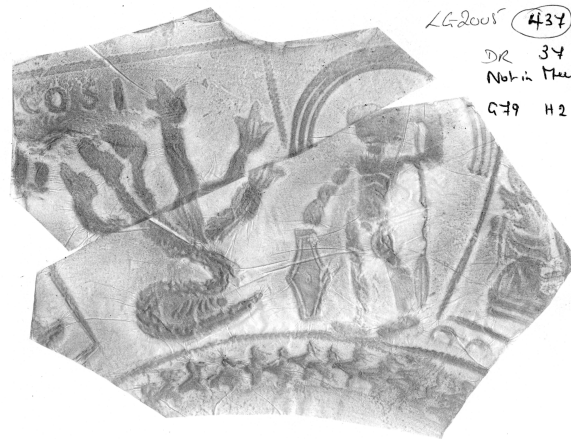


Figure 2. Rubbing of a decorated Samian (Terra Sigillata) vessel, made in La Graufesenque, 100-120 AD, decorated by the potter L. Cosius. [Geoffrey Dannell and Allard W. Mees, CC BY 4.0, via Wikimedia Commons]

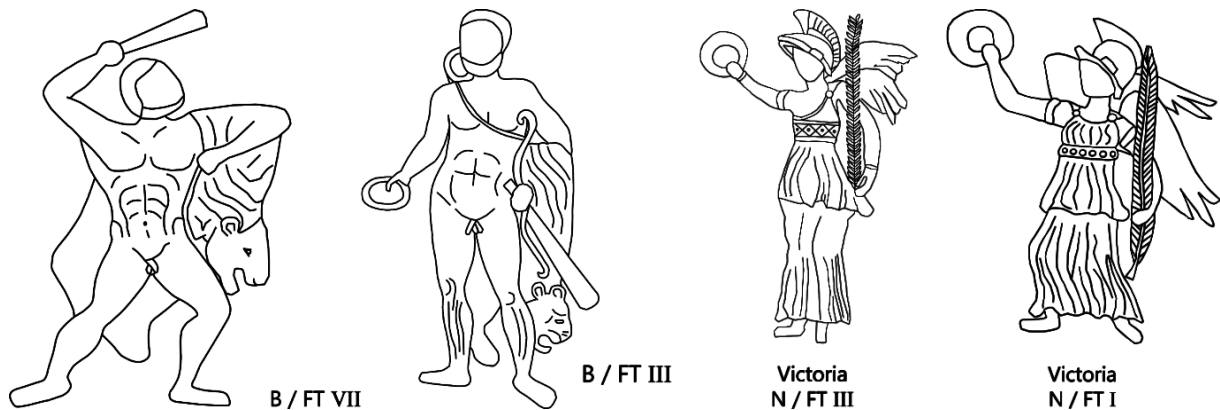


Figure 3. Examples for representations of Hercules (B) [10, p.702] and Victoria (N) [10, p.719] on Roman terra sigillata, reproduced with permission from Sophie zu Löwenstein, published in [10] and [11].

The previous practice of recording ARS pottery and its decoration involved photographs and drawings in analogue publications [4] (Figures 2 , 3). However, for reproduction and comparison of the appliqué and their figure types, this method is not entirely suitable, for it does not allow for an accurate recording of the plasticity of the objects. The lack of standardised assignments for figure types also hampers the comparison. For the current research, structured, machine-readable data enriched with specialised information is required to reflect the complexity of the object's information fabric. In a recent paper, a workflow for innovative, comprehensible and sustainable FAIR data access for this material has been developed [9]. The results and methods are being taken up and further developed in N4O and some overlapping consortia such as NFDI4Culture (N4C).



Figure 4. Data acquisition. [CC BY-SA 4.0 ARS3D project / i3mainz / LEIZA, published in [5]]

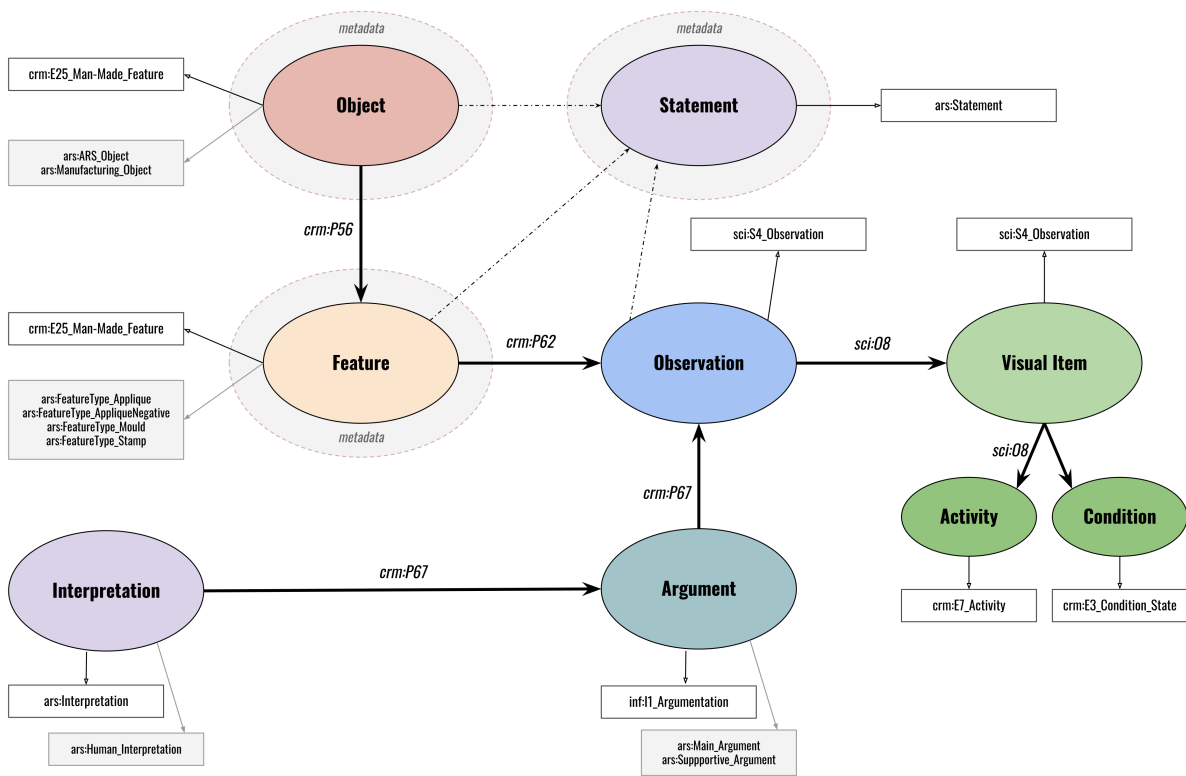


Figure 5. (Schematic illustration of the ARS3D semantic modelling approach. Objects, features, observations and interpretations are modelled based on CIDOC CRM and its extensions (Observation Modelling Scheme, containing visual items, activities and conditions). [Florian Thiery, CC BY 4.0, published in [9]]

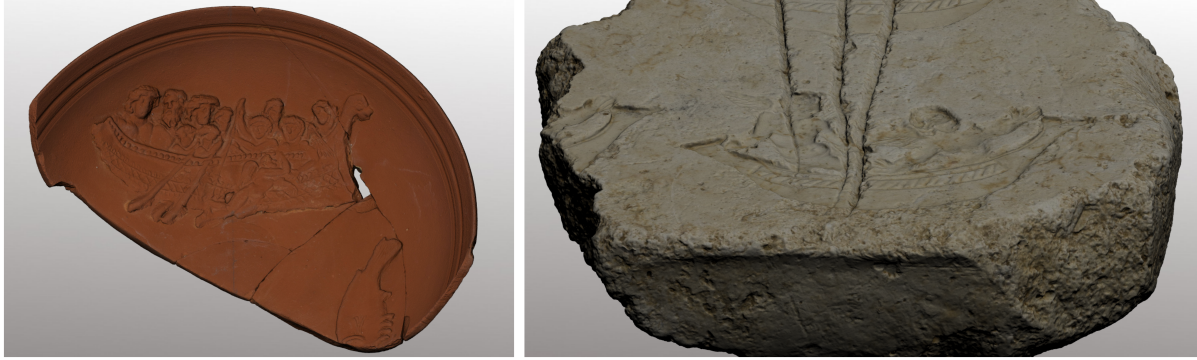


Figure 6. (Left) Appliqués applied to vessel O.41260; (right) patrix impression in a mould in order to create the appliqué on O.41418 [ARS3D Project/i3mainz/RGZM, CC BY-SA 4.0, via Wikimedia Commons]

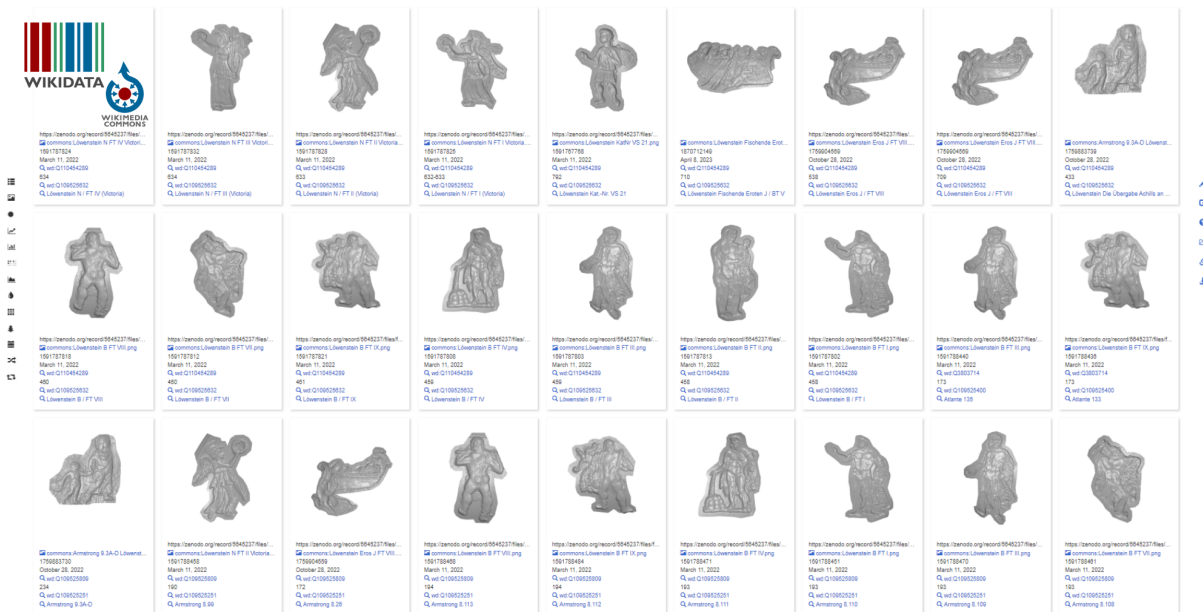


Figure 7. Wikidata query service: Iconographic items of the ars3d project. [Wikimedia Community, CCO, Public Domain]

The FAIRification workflow as a use-case consists of three steps to enable digital research, e.g. comparisons of appliqués on ARS [5], [9]:

1. Geometric capturing [5], [9], [12] by standardised 3D digitalisation workflows, e.g. using a structured light projection scanner and a camera to compute a textured model (Figure 4); this is addressed by N4O TA1 for Documentation and N4C. This process is accompanied by documenting the 3D capturing and processing metadata in an ontology [13].
2. Semantic (meta) data modelling (Figure 5) by community standards, ontologies (e.g., CIDOC CRM, PROV-O) and controlled vocabularies [9]. This enables annotating archaeological features in 3D models and referencing (e.g. the motifs on the ARS vessels) with links to IconClass, Getty AAT and Wikidata. These issues are addressed in N4O TA2 for Collecting and in N4C where similarities to the Kompakkt 3D-Web-Viewer exist, allowing for semantic enrichment of 3D-models by annotating them with Wikibase links.

3. The visualisation and publication of the semantically enhanced 3D models by web tools such as 3DHOP (Figure 6) and Linked Open Data [5], [9], [11], [14]. To enable the engagement of Citizen Scientists, this data is also integrated and published in Wikidata [15], [16] (Figure 7), as well as in iDAI.objects, to provide long-term availability. These tasks are addressed within N4O TA5 for Storage, Access and Dissemination, as well as in N4C.

4 From the Use Case to the NFDI

This exemplary workflow can be easily adapted to (1) other object groups and materials, e.g. semantic modelling of iconography in numismatics, (2) other contexts, e.g. highly precise documentation of excavations and surveys by using controlled vocabularies, (3) other data qualification methods, e.g. authority data enrichments for archaeobotanical samples (Arbodat), (4) enrichment of object data by Wikidata, as well as the integration of data into the Community Hub (semantic up-/downlift), (5) provision and modelling of material groups with ethical issues by applying the CARE principles, and (6) usage of the N4O infrastructure, e.g. iDAI.world, DANTE or archaeology.link. This workflow for and approaches to achieving and providing FAIR data suit the implementation in other NFDI domains, especially in the fields of controlled vocabularies such as ontologies and thesauri, as well as metadata and community standards.

Data availability statement

Data described in this paper is available OSF (DOI: 10.17605/OSF.IO/6HJ7G), Zenodo (<https://zenodo.org/communities/ars3d> accessed on 26 March 2023) and on Wikidata. Especially: the comparison data (DOI: 10.5281/zenodo.5647827), comparison scripts (DOI: 10.5281/zenodo.5647864), graph data (DOI: 10.5281/zenodo.5642750), ontology (DOI: 10.5281/zenodo.5642891), images of features (DOI: 10.5281/zenodo.5645236), comparisons (DOI: 10.5281/zenodo.5645253), the source code of the web application (DOI: 10.17605/OSF.IO/P5TKW) and of the API (DOI: 10.17605/OSF.IO/WRJ7K).

Competing interests

The authors declare that they have no competing interests.

Funding

The African red slip ware in digital form – 3D documentation for the multi-perspective analysis of a central object category of the late antiquity period Project (ARS3D) was funded by the Federal Ministry of Education and Research Germany (BMBF), Förderkennzeichen: BMBF-01UG1888AX, BMBF-01UG1888BX.

Acknowledgements

This paper is written with the help of the whole NFDI4Objects consortium. We would like to thank the Executive Director Christin Keller, and Coordination Office member Henrike Backhaus, and all Co-Spokespersons. We would also like to thank Vladimir Stolba for English support.

References

- [1] D. Bibby, K.-C. Bruhn, F. Dürhkohp, C. Eckmann, U. Himmelmann, B. Höke, C. Keller, M. Lang, A. Mees, S. Metz, M. Renz, P. v. Rummel, F. Schäfer, H. Senst, J. Sessing, T. Stöllner, F. Thiery, B. Weisser, and D. Wintergrün, "Digitales Forschungsdatenmanagement in der Archäologie und die Initiative NFDI4Objects", de, *Blickpunkt Archäologie*, vol. 2, pp. 150–163, 2021. DOI: [10.5281/zenodo.5823867](https://doi.org/10.5281/zenodo.5823867).
- [2] J. Hayes, *Late Roman Pottery*. London, 1972.
- [3] M. Bonifay, *Études sur la céramique romaine tardive d'Afrique*, ser. British Archaeological Reports International Series 1301. Archaeopress, 2004, p. 525. [Online]. Available: <https://shs.hal.science/halshs-01956529>.
- [4] M. Armstrong, "A thesaurus of applied motives on african red slip ware", English, PhD thesis, New York University, 1993.
- [5] F. Thiery, L. Raddatz, and F. Boochs, "Close to the Original-Erfassung archäologischer Objekte und ihre webbasierte semantisch modellierte Bereitstellung zur fachwissenschaftlichen Analyse", in *Photogrammetrie - Laserscanning - Optische 3D-Messtechnik: Beiträge der Oldenburger 3D-Tage 2022*. Berlin: Wichmann, 2022, pp. 56–67, ISBN: 978-3-87907-726-7.
- [6] M. Mackensen, *Relief- und stempelverzierte nordafrikanische Sigillata des späten 2. bis 6. Jahrhunderts: römisches Tafelgeschirr der Sammlung K. Wilhelm*, ger, ser. Münchner Beiträge zur provinzialrömischen Archäologie Band 8, Teil 1. Wiesbaden: Reichert Verlag, 2019, ISBN: 978-3-95490-413-6.
- [7] J. Garbsch, "Spätantike sigillata-tabletts", *Rei Cretariae Romanae Fautores*, vol. Zur spätantiken Keramik aus Nordafrika, Supplementa, Vol. 5, no. Supplementa, Vol. 5, 1980.
- [8] M. Armstrong, "The köln römisch-germanisches museum study collection of african red slip ware", *Kölner Jahrbuch für Vor- und Frühgeschichte* 24 (1991), vol. 24, pp. 412–475, 1991.
- [9] F. Thiery, J. Veller, L. Raddatz, L. Rokohl, F. Boochs, and A. W. Mees, "A Semi-Automatic Semantic-Model-Based Comparison Workflow for Archaeological Features on Roman Ceramics", en, *ISPRS International Journal of Geo-Information*, vol. 12, no. 4, p. 167, Apr. 2023, ISSN: 2220-9964. DOI: [10.3390/ijgi12040167](https://doi.org/10.3390/ijgi12040167). [Online]. Available: <https://www.mdpi.com/2220-9964/12/4/167> (visited on 04/26/2023).
- [10] S. von Löwenstein, *Mythologische Darstellungen auf Gebrauchsgegenständen der Spätantike: die appliken- und reliefverzierte Sigillata C3/C4*, ser. Kölner Jahrbuch 48. 2015.
- [11] S. C. Schmidt, F. Thiery, and M. Trognitz, "Practices of linked open data in archaeology and their realisation in wikidata", en, *Digital*, vol. 2, no. 3, pp. 333–364, Sep. 2022, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2673-6470. DOI: [10.3390/digital2030019](https://doi.org/10.3390/digital2030019). [Online]. Available: <https://www.mdpi.com/2673-6470/2/3/19> (visited on 06/22/2022).
- [12] C. Justus, P. Atorf, and F. Boochs, "Gewinnung realitätsnaher virtueller Modelle als Grundlage für die Erkennung von ähnlichkeiten", English, in *Photogrammetrie - Laserscanning - Optische 3D-Messtechnik: Beiträge der Oldenburger 3D-Tage 2019*. Berlin: Wichmann, 2019, pp. 250–258, ISBN: 978-3-87907-660-4.
- [13] T. Homburg, A. Cramer, L. Raddatz, and H. Mara, "Metadata schema and ontology for capturing and processing of 3D cultural heritage objects", *Heritage Science*, vol. 9, no. 1, p. 91, 2021, ISSN: 2050-7445. DOI: [10.1186/s40494-021-00561-w](https://doi.org/10.1186/s40494-021-00561-w). [Online]. Available: <https://doi.org/10.1186/s40494-021-00561-w> (visited on 03/24/2022).
- [14] F. Thiery and L. Rokohl, "Linked open african red slip ware", *Squirrel Papers*, vol. 3, no. 1, No.3, Nov. 2021. DOI: [10.5281/zenodo.5722941](https://doi.org/10.5281/zenodo.5722941). (visited on 02/24/2022).

- [15] F. Thiery, A. Mees, and L. Rokohl. (2022). Wikidata: Wiki-project african red slip ware digital, [Online]. Available: https://www.wikidata.org/wiki/Wikidata:WikiProject_African_Red_Slip_Ware_Digital (visited on 06/21/2022).
- [16] Wikidata Community. (2023). Wikidata query service: Iconographic items of the ars3d project, [Online]. Available: <https://w.wiki/4paq> (visited on 04/26/2023).

WissKI

A Virtual Research Environment based on Drupal

Mark Fichtner¹[\[https://orcid.org/0000-0001-5597-4222\]](https://orcid.org/0000-0001-5597-4222), Robert Nasarek²[\[https://orcid.org/0000-0003-1163-8504\]](https://orcid.org/0000-0003-1163-8504), and Tom Wiesing³[\[https://orcid.org/0009-0002-7392-0556\]](https://orcid.org/0009-0002-7392-0556)

¹ Germanisches Nationalmuseum, Germany

² Germanisches Nationalmuseum, Germany

³ Friedrich-Alexander Universität Erlangen-Nürnberg, Germany

Abstract. WissKI is a free and open source virtual research environment based on the free and open source content management system Drupal. It features everything that the content management system provides while using a triplestore for authoritative data storage. Thus changes can be made in the triplestore and they are directly reflected by the system. Furthermore WissKI provides all features that are necessary for a full Linked Open Data Semantic Web platform.

Keywords: WissKI, Linked Open Data, Semantic Web, Virtual Research Environments

Introduction

Increasing digitisation and dealing with the Semantic Web are a central challenge of the future for the traditional humanities. In this context, the digital medium not only supports research, but also poses challenges in terms of the change in methods, the speed of development and the openness of research data. In recent years, virtual research environments have increasingly established themselves as research platforms in research projects. A constant point of criticism from the user's point of view is that dealing with virtual research environments usually requires a certain affinity to the digital medium and a greater period of familiarisation than was necessary for the traditional collection of data in the analogue domain. The central approach of virtual research environments, however, is to maintain a balance between complexity or a large number of application possibilities on the one hand and a low entry threshold with simple usability of the data on the other. The following is an example of the implementation of such an approach from the WissKI research project funded by the DFG.

WissKI Project

As part of the three-year and subsequently two-year DFG-funded research project "Scientific Communication Infrastructure" (WissKI, [1]), the virtual research infrastructure of the same name "WissKI" was created on the basis of the open-source content management system Drupal. In cooperation between the Germanisches Nationalmuseum Nürnberg (GNM), the Zoologisches Forschungsmuseum Alexander Koenig in Bonn (ZFMK) and the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), a software infrastructure was developed from the requirements of cooperative research in the field of cultural heritage and its documentation in digital media. The focus of the system is not only the simple provision and open availability of source materials - structured texts, graphics, images, video, audio - and metadata in digital

form, but also collaboration on the basis of semantic web principles. By leaving the typical features of common content management systems untouched, the system has detailed user control with rights management and is able to manage web content such as websites, forums and wikis and present them online.

The software is published free of charge as open source on the Drupal module website [1] and can therefore be used, reused and extended accordingly.

Integration in Drupal

The current version of the WissKI software is based on the currently most modern version of the content management system Drupal - Drupal 10 [2]. The functionality of Drupal can be extended and modified by third-party modules. Accordingly, the WissKI system is a set of modules divided into logical units, each of which brings encapsulated functionality to the system. The modules are based on all core functionalities of Drupal, so that all common features of Drupal, such as user control with detailed rights management or the creation of web pages, are retained.

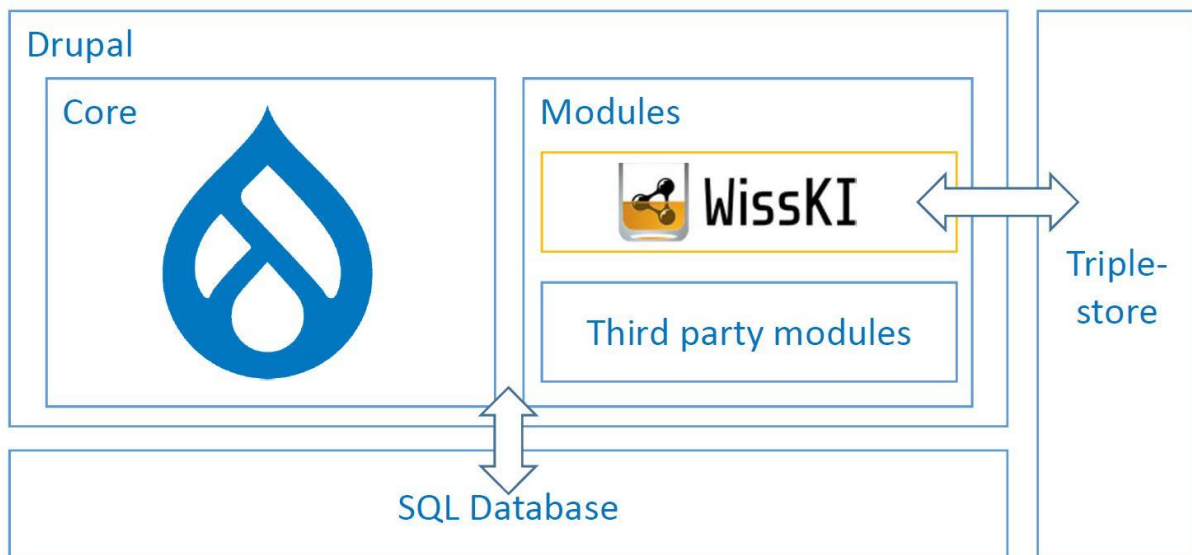


Figure 1. WissKI integration in Drupal

Semantic modelling with Drupal and WissKI

The focus of the WissKI project was to combine the implementation of the common data storage of the Semantic Web in the form of triples with the functionality of Drupal. An approach

such as Semantic Mediawiki [3] would be obvious: All individuals correspond to their own web-sites and all relations correspond to data fields. If one tries to implement this directly with Drupal, individuals in Drupal correspond to entities, concepts to bundles and relations to fields.

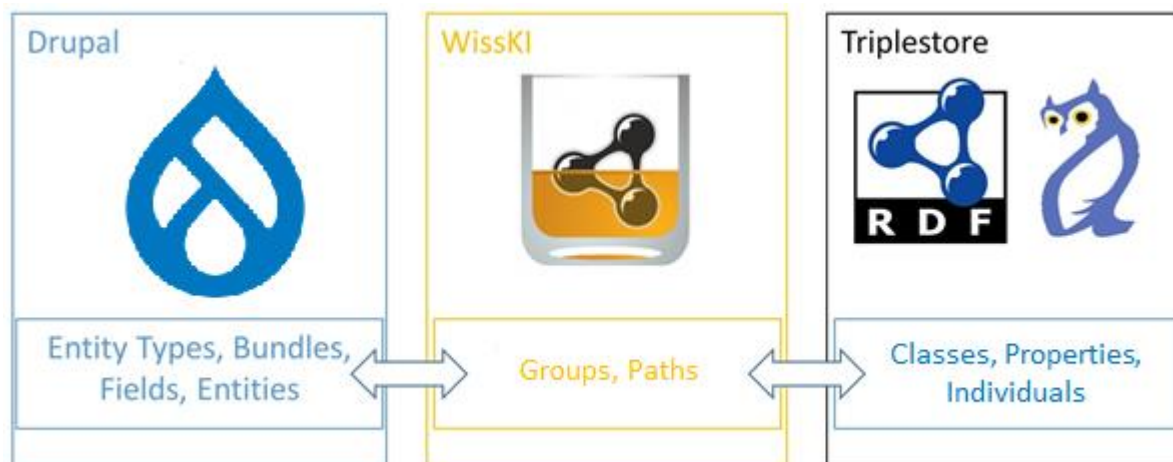


Figure 2. Translation of Drupal contents to the Triplestore via WissKI

Due to the strong requirements of cultural heritage ontologies like the CIDOC Conceptual Reference Model (CIDOC CRM, ISO 21127, [4]) for a complex data graph, WissKI implements an even more complex approach. The integral component of the system is the so-called "Pathbuilder". It supports the system administrator in creating "paths" through the ontology. A path is a concatenation of n concepts and $n-1$ relations between the concepts. When storing data using a path, an individual is first created for each concept. Then the individuals created are connected to each other by the relations according to the specifications of the path. At the end of such a path in WissKI there is always a relation to a primitive data type in which the actual input is stored. To store several inputs via the same subject, paths can be combined into so-called groups. The group defines the common portion of all paths belonging to it. The trivial case (groups and paths of length 1) corresponds directly to the implementation of OWL described above. However, it additionally allows the implementation of more complex modeling as prescribed by the CIDOC CRM. For the implementation in Drupal, the Pathbuilder forms an intermediate layer between the Triplestore with the data stored there in triples on the one hand and Drupal with the data storage based on entity types, bundles, entities and fields on the other. It creates a mapping from groups and subgroups in Pathbuilder to bundles and referenced bundles in Drupal as well as from paths to data fields. This mechanism hides the full complexity of the Semantic Web approach as well as the CIDOC CRM from the actual user, who only has to fill in forms.

The system gives the possibility to load any ontology based on OWL. When creating paths based on the ontology, the system can support the administrator by calculating the possible concepts and relations for each step (based on domain and range).

Conclusion

The WissKI system offers the administrator many options for support and the user the possibility to store his or her data on the basis of common semantic web technology. However, the complexity of this process is not reduced, but shifted from the user to the administrator. The administrator needs a sound knowledge of Drupal, the Semantic Web and the standards on which it is based, as well as the CIDOC CRM capture standard. At the same time, he is responsible for implementing the intended semantics of the users in the database in the form of masks and fields. With this approach, the WissKI system tries to keep the entry hurdle for the user low and at the same time offer the administrator all the possibilities of the Semantic Web.

References:

1. "ABOUT WISSKI | wiss-ki.eu". <https://wiss-ki.eu/> (26.04.2023)
2. "WissKI | Drupal.org". <https://www.drupal.org/project/wisski> (26.04.2023)
3. "Drupal – Open Source CMS | Drupal.org". <https://www.drupal.org/> (26.04.2023)
4. "semantic-mediawiki.org". https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki (26.04.2023)
5. "Home | CIDOC CRM". <https://cidoc-crm.org/> (26.04.2023)

MetaBelgica Project

A Linked Data Infrastructure Between Federal Scientific Institutes in Belgium

Sven Lieber¹[\[https://orcid.org/0000-0002-7304-3787\]](https://orcid.org/0000-0002-7304-3787), Ann Van Camp¹[\[https://orcid.org/0000-0002-1915-5956\]](https://orcid.org/0000-0002-1915-5956),
Dieter De Witte²⁵[\[https://orcid.org/0000-0001-8480-5719\]](https://orcid.org/0000-0001-8480-5719), Eva Coudyzer³[\[https://orcid.org/0000-0002-5985-7231\]](https://orcid.org/0000-0002-5985-7231),
Erik Buelinckx³[\[https://orcid.org/0000-0003-1831-158X\]](https://orcid.org/0000-0003-1831-158X), Els Angenon⁴[\[https://orcid.org/0000-0002-8888-9662\]](https://orcid.org/0000-0002-8888-9662),
Hannes Lowagie¹[\[https://orcid.org/0000-0002-0671-3568\]](https://orcid.org/0000-0002-0671-3568), Julie Birkholz¹⁵[\[https://orcid.org/0000-0003-1193-0847\]](https://orcid.org/0000-0003-1193-0847),
Karine Lasaracina²[\[https://orcid.org/0009-0006-1732-6607\]](https://orcid.org/0009-0006-1732-6607)

¹Royal Library of Belgium (KBR), Brussels, Belgium

²Royal Museums of Fine Arts (RMFAB), Brussels, Belgium

³Royal Institute for Cultural Heritage (KIK-IRPA), Brussels, Belgium

⁴Royal Museums of Art and History (RMAH), Brussels, Belgium

⁵Ghent University, Ghent, Belgium

Abstract:

Keywords: FAIR Data, Linked Data, RDF, Wikibase, GLAM, Belgium

1 Introduction

Trustworthy metadata about entities related to cultural heritage is important to correctly identify bibliographic metadata, attribute works, identify works of public domain (based on contributors' date of death), or to support Named Entity Linking (NEL) for digitised documents. However, in Belgium such data is currently dispersed between numerous institutions, modelled with different ontologies, represented in different formats and curated in different languages. GLAM institutions would profit from data about Belgian entities to correctly annotate their collections. Furthermore, this would facilitate research in Belgium and abroad. The current situation does not only complicate data exchange in a national and international setting, but also leads to duplicate data curation efforts and data quality issues, negatively impacting the users' experience.

This paper introduces the *MetaBelgica* project, coordinated by KBR - The Royal Library of Belgium with the Royal Museums of Fine Arts, the Royal Institute for Cultural Heritage, and the Royal Museums of Art and History, to improve the status quo. The aforementioned *Federal Scientific Institutes (FSIs)*, belonging to the *Belgian Science Policy Office (BELSPO)*, joined forces to develop a shared Linked Data platform for managing shared entities. This platform, based on Wikibase (the technology behind Wikidata [1]), aims to ensure FAIR data about *persons, organisations, time/events* and *locations* related to Belgian cultural heritage. We will integrate data about millions of

Belgian entities from the four participating FSIs into a Wikibase instance and make it accessible by using Persistent Identifiers. This will not only professionalize our own data management and improve data quality, but according to received project support letters, also impact other regional, national and European institutions.

The aim of this paper is to introduce the project and its methodology as to to obtain early feedback and exchange information with the Research Data Infrastructure community. The *MetaBelgica* project will kick-off in 2023 and will run until the end of 2026. Section 2 presents the methodology - anchored in the state of the art - and section 3 concludes this paper.

2 Methodology

The proposed methodology aims to create a sustainable entity management system for Belgian entities to serve different stakeholders, and therefore covers (i) the integration of existing FSI data, (ii) a platform to manage the data in the long term, and (iii) providing data access to different stakeholders. The proposed methodology consists of five pillars that are translated into iteratively executed work packages and tasks, see fig. 1.

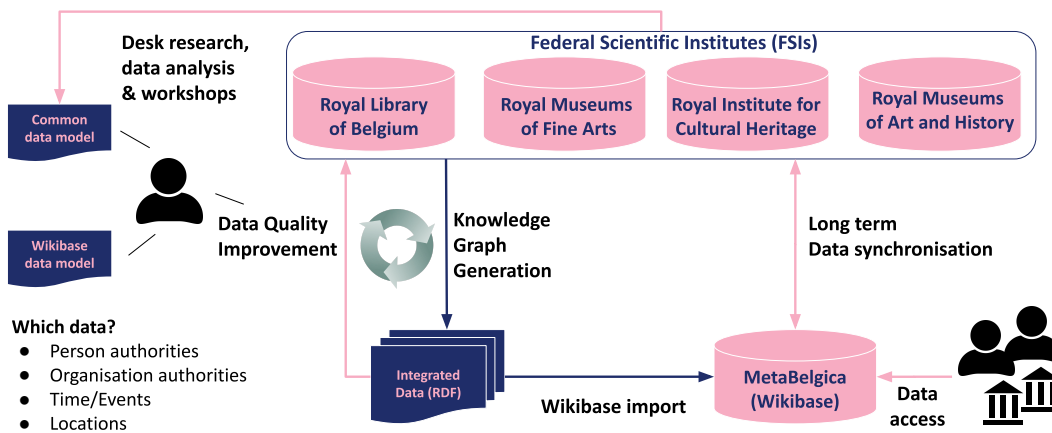


Figure 1. An illustration of the MetaBelgica setup: Data from the four FSIs is integrated according to a common RDF data model. Based on an additional mapping to a Wikibase data model, the data will be ingested into a Wikibase instance from which it can be collaboratively curated by the FSIs and from where it will be accessible.

Review current practices and data, and develop a common data model By using desk research and workshops, we will review the state of the art to represent the selected entities in an interoperable fashion. For example by using regional, national and international standards such as OSLO [2], FedVoc [3], or the Europeana Data Model [4]. Furthermore, we will analyse the existing data and divide it into different quality categories to ease further integration steps and link suitable concepts of international taxonomies.

A divide and conquer approach to integrate data in an iterative fashion By following state of the art knowledge engineering activities, we will iteratively integrate FSI data, e.g. with declarative RML mapping rules [5] for Knowledge Graph Construction of heterogeneous data. We will determine different integration strategies based on the identified quality categories, i.e. semi-automatic integration for high quality RDF data based on identifiers (with a human-in-the-loop [6]) or a manual curation for lower quality data.

Collaborative entity management software To enable coordinated data curation between FSIs, we will set up a collaborative Wikibase system that can be used for data curation in the long term and integration of low-quality data via reconciliation during the project. We develop a Wikibase data model and investigate suitable data ingestion methods such as the recently announced Wikibase REST API or other automatic means [7]. The open source software Wikibase is already tested in other projects [8] and comparable international initiatives such as (i) the *French National Entities File* [9] by the French national library BnF and the Bibliographic Agency for Higher Education (ABES), (ii) the *Integrated Authority File (GND)* in Austria, Germany and Switzerland [10], and (iii) The *Shared Authority File* in Luxembourg [11].

Long-term and sustainable data curation platform Longevity of the platform is a core feature of this project. Besides a stronger collaboration between the FSIs, we want to ensure the sustainability of *MetaBelgica* by integrating its operation into the functioning of each institution. We will set up organisational structures and provide internal trainings to ensure coordinated data curation with the new platform. Furthermore, we will implement technical components to ensure data synchronisation in the long term. Therefore we also have chosen for Wikibase: technical personnel is rare in the GLAM field, but data in a Wikibase can be maintained by non-technical personnel such as librarians or curators.

Data access via open license to increase impact We will ensure that the dispersed and heterogeneous data are presented to different users in a uniform and persistent way as open data. This includes surveying relevant stakeholders to provide appropriate FAIR data services. For these activities, we will follow the BELSPO Open Data directive "as open as possible and as close as necessary".

3 Conclusion

MetaBelgica has the aim to create FAIR entities of Belgian cultural heritage for worldwide use. The primary stakeholder group is the scientific community, but the platform can also be of use for society in general. Targeted users are scientific (researchers), cultural (GLAM-professionals and a broad public), educational (teachers and academics), technical (data aggregators) and economic (publishing professionals and creators) stakeholders.

With respect to the creation and operation of the presented platform we anticipate different challenges related to legal, organisational and technical interoperability. Challenges, that other Research Data Infrastructure projects (of other domains) likely encounter as well and for which best practices should be established.

Competing interests

The authors declare that they have no competing interests.

Funding

The presented project is funded by the Belgian Science Policy Office (BELSPO) in an Impulse action call to fostering the development of emerging research infrastructures within federal research institutions (INFRA-FED).

Acknowledgements

The authors like to thank the various stakeholders that supported the project proposal with a letter of intent.

References

- [1] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, Sep. 2014, ISSN: 0001-0782. DOI: [10.1145/2629489](https://doi.org/10.1145/2629489).
- [2] R. Buyle, L. De Vocht, M. Van Compernelle, et al., "OSLO: Open standards for linked organizations," en, in *Proceedings of the International Conference on Electronic Governance and Open Society: Challenges in Eurasia*, St. Petersburg Russia: ACM, Nov. 2016, pp. 126–134, ISBN: 9781450348591. DOI: [10.1145/3014087.3014096](https://doi.org/10.1145/3014087.3014096).
- [3] FPS BOSA, *Federal Vocabularies*, English, 2019. [Online]. Available: <https://www.belgif.be/page/specification/fedservicesplatform.en.html>.
- [4] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, and H. Van de Sompel, "The Europeana Data Model (EDM)," in *World Library and Information Congress: 76th IFLA general conference and assembly*, vol. 10, 2010, p. 15.
- [5] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle, "RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data," in *Proceedings of the 7th Workshop on Linked Data on the Web*, C. Bizer, T. Heath, S. Auer, and T. Berners-Lee, Eds., ser. CEUR Workshop Proceedings, vol. 1184, CEUR-WS.org, 2014. [Online]. Available: http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.
- [6] Lieber, Sven, Van Camp, Ann, and Lowagie, Hannes, *A LITL more quality: Improving the correctness and completeness of library catalogs with a Librarian-In-The-Loop Linked Data Workflow*, en, Nov. 2022. DOI: [10.5281/ZENODO.7372985](https://doi.org/10.5281/ZENODO.7372985).
- [7] R. Shigapov, J. Mechnich, and I. Schumm, "RaiseWikibase: Fast Inserts into the BERD Instance," en, in *The Semantic Web: ESWC 2021 Satellite Events*, R. Verborgh, A. Dimou, A. Hogan, et al., Eds., vol. 12739, Cham: Springer International Publishing, 2021, pp. 60–64, ISBN: 9783030804176. DOI: [10.1007/978-3-030-80418-3_11](https://doi.org/10.1007/978-3-030-80418-3_11).
- [8] J. Godby, K. Smith-Yoshimura, B. Washburn, et al., "Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage," DOI: [10.25333/faq3-ax08](https://doi.org/10.25333/faq3-ax08).
- [9] B. Bober and A. Angjeli, *Assessing Wikibase as the core for the French National Entities File (FNE)*, English, Berlin, Oct. 2019. [Online]. Available: https://upload.wikimedia.org/wikipedia/commons/1/16/Wikibase_for_FNE.pdf.
- [10] B. Fischer and S. Hartmann, *GND meets wikibase*, English, Jul. 2020. [Online]. Available: <https://www.youtube.com/watch?v=3vJrsUQIbV4>.
- [11] J. E. Labra Gayo, M. Pfeiffer, A. Waagmeester, et al., *Representing the Luxembourg Shared Authority File based on CIDOC-CRM in Wikibase*, Semantic Web in Libraries Conference 2021, Dec. 3, 2021. [Online]. Available: <https://www.youtube.com/watch?v=MDjyiYr0WJQ> (visited on 04/21/2023).

Sociodemographic variables in surveys

Increasing research potential through output harmonization

Silke Schneider¹ and Lennart Palm¹

¹ GESIS - Leibniz-Institute for the Social Sciences

Measuring people's sociodemographic characteristics over time and in different contexts or studies is one of the keys of quantifying social inequalities and social change within any society. While different respondent-based studies usually focus on different topics, almost all of them collect data on some key features defining respondent's backgrounds and living situations.

Summarizing the measurement of sociodemographic characteristics however, the German survey landscape could be characterized as "same but different": studies measure the same concepts yet differ in their specific approach, despite the fact that the "Demographische Standards" (Hoffmeyer-Zlotnik et al., 2016) provide recommended questionnaire items since the early 1990s. This has several problematic implications: Firstly, it becomes quite difficult to combine knowledge about groups of respondents with the same sociodemographic characteristics across studies on a macro level, and thus to systematically accumulate knowledge on social inequalities and social change. Secondly, combining different data sets, e.g. in order to be able to analyze small groups or rare phenomena, can become a laborious effort which might still produce questionable results, since key caveats can be overlooked in survey documentation. Combining supposedly basic categories such as marital status across studies can already make simplification and therefore loss of information necessary, as there are multiple ways of measuring marital status as Table 1 shows.

Table 1: Overview of (proposed) measurements of marital status in selected studies by Schneider et al (2022)¹

Legally possible categories of marital status in Germany	Categories in Dem. Standards, ALLBUS, G Panel, FReDA, GIP	Categories in MZ, SOEP	Categories in GLES, Covid-19-U	Categories in GEDA, NEPS	Categories in Best_FDM ² (minimal)
Married, living with spouse	X	X*	X		
Registered partnership, living with partner	X	X**	X	X	X
Married, not living with spouse	X	X*	X	X	
Registered partnership, not living with partner	X	X**	X		
Divorced	X	X	X	X	X
Registered partnership, dissolved	X	X			
Widowed	X	X	X	X	X
Registered partner passed	X	X			
Single	X	X	X	X	X

ALLBUS: Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. <https://www.gesis.org/en/allbus/allbus-home>

Best_FDM: connecting empirical social science research in societal crises. <https://wzb.eu/en/research/trans-sectoral-research/konsortswd/projects/bestfdm-vernetzung-empirischer-sozialwissenschaftlicher-forschung-in-gesellschaftlichen-krisen>

Covid-19-U: COVID-19 and Inequality - Survey Program. <https://www.exc.uni-konstanz.de/en/inequality/research/covid-19-and-inequality-surveys-program/>

Dem. Standards: Demographic Standards. https://www.statistischebibliothek.de/mir/receive/DE-Monografie_mods_00003695

FReDA: The German Family Demographic Panel Study. <https://www.freda-panel.de/FReDA/EN/Startseite.html>

G Panel: GESIS Panel. <https://www.gesis.org/en/gesis-panel/gesis-panel-home>

GIP: German Internet Panel. <https://www.uni-mannheim.de/en/gip/>

GLES: German Longitudinal Election Study. <https://www.gesis.org/wahlen/gles>

MZ: Microcensus. <https://www.gesis.org/en/missy/metadata/MZ/>

NEPS: National Educational Panel Study. <https://www.neps-data.de/Mainpage>

SOEP: Socio-Economic Panel. https://www.diw.de/de/diw_01.c.412809.de/sozio-oeconomisches_panel_soep.html

**, **: MZ and SOEP don't differentiate between spouses/partners living together or not in their first marital status item, yet do so later on. Same number of stars indicate same category in marital status item.*

It would be desirable if harmonization of basic sociodemographic concepts was done once and for all to use. To facilitate output harmonisation of socio-demographic variables, KonsortSWD develops proposals for standard variables for selected socio-demographic attributes.

¹ Schneider et al (2022): <https://zenodo.org/record/6810973#.ZBhhUoSZOU>

² These are recent recommendations for questionnaire items by RatSWD, which were therefore not included in the review by Schneider et al (2022). For more details see: <https://zenodo.org/record/6810973#.ZBhhUoSZOU>

While our approach was based on the German survey data landscape, we leaned upon international standards such as ISCED for education or the EU's standardised key social variables for household net income or main activity status, to ensure international compatibility. To provide for the quality and usefulness of the proposed standard variables, we used three methods: Firstly, before developing our proposals, we reviewed existing survey instruments in several of Germany's leading studies (<https://doi.org/10.5281/zenodo.6810973>). Secondly, the proposals were discussed in a virtual roundtable meeting with researchers, study representatives and data users, as well as bilaterally with individual experts. Thirdly, based upon a multiple linear regression analysis approach, we validated our proposals both in a data-driven and a theory-driven way, using a broad set of up to 190 potential outcome variables. Based on these validation results and the feedback gained, we have refined our proposals.

Ideally, our proposals would be published within scientific use files of individual studies for easy access and usage.

Additionally, we hope our proposals could help implement an understanding for time constant variables and their specific needs in data collection when unavoidable changes to questionnaire items, which are repeatedly necessary to keep up with societal changes and respondent's life's and understanding, happen. Another benefit of this path of harmonisation could also be greater independency from survey mode switches due to technological changes and technological use patterns within societies if data needs of our proposals are considered when mode switches occur.

In this talk, we will showcase the standard variables for 3 socio-demographic attributes, namely education, marital status and main activity status to collect final feedback. Our proposals will be published in autumn 2023.

Competing interests

The authors declare that they have no conflict of interest.

References

1. Hoffmeyer-Zlotnik, J. H. P., Beckmann, K., Glemser, A., Heckel, C., von der Heyde, C., Schneider, S. L., Hanefeld, U., Herter-Eschweiler, R., & Kühnen, C. (2016). *Demographische Standards, Ausgabe 2016: Eine gemeinsame Empfehlung des ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V., der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V. (ASI) und des Statistischen Bundesamtes*. Statistisches Bundesamt. https://www.statistischebibliothek.de/mir/servlets/MCRFileNo-deServlet/DEMonografie_derivate_00001549/Band17_DemographischeStandards1030817169004.pdf

Stamp - Standardized Data Management Plan for Educational Research

An Approach to Improve Cross-Disciplinary Harmonization of Research Data Management

Sebastian Netscher¹[\[https://orcid.org/0000-0002-2784-6968\]](https://orcid.org/0000-0002-2784-6968), Alexia Meyermann², Julia Künstler-Sment², and Lisa Pegelow³[\[https://orcid.org/0000-0003-4148-6978\]](https://orcid.org/0000-0003-4148-6978)

¹ GESIS - Leibniz-Institute for the Social Sciences

² DIPF | Leibniz Institute for Research and Information in Education

³ Institute for Educational Quality Improvement

Keywords: Research Data Management, Standardized Data Management Plan, Harmonization of RDM

Introduction

While there is a strong tendency towards a harmonized, cross-disciplinary research data management (RDM) (Netscher et al., 2022), researchers require more guidelines and examples, tailored to their research discipline (Grootveld et al., 2018). Therefore, Science Europe (2018: 9) proposes developing so-called domain data protocols (DDP), “a ‘model DMP’ for a given domain or community”. Based on this concept, the project *Domain Data Protocols for Educational Research*¹ designed the *Stamp - Standardized Data Management Plan for Educational Research* (DDP-Bildung and German Network of Educational Research Data, 2023).

Although the Stamp was designed to support researchers in educational research, we expect that RDM is rather a matter of the data processed, the methods employed, and the content of data, than of a particular research discipline or community, such as educational research. To discuss this expectation and the usability of the Stamp outside educational research, we organized various workshops with representatives from other research disciplines. In our talk at the *CoRDI 2023*, we will introduce the Stamp, recap findings of two of the workshops, introduce the next steps to examine the useability of the Stamp outside educational research, and draw some conclusions on how to adapt the Stamp to other disciplines and how it fosters a harmonized, cross-disciplinary RDM.

Stamp - Standardized Data Management Plan for Educational Research

The Stamp composes of a so-called *basic module* and eight *content modules*, illustrated in Figure 1. The basic module structures RDM and provides information on, e.g., the project and the data processed. The eight content modules cover different topics of RDM, such as research ethics, data documentation and traceability, or data sharing. Each content modules consists

¹ The project DDP-Bildung was funded by the German Federal Ministry of Education and Research (grant number: 16QK01).

of a minimal condition, a short statement on how to manage data in the context of the respective module to ensure processing shareable data, according to the FAIR Data Principles (Wilkinson et al., 2016).

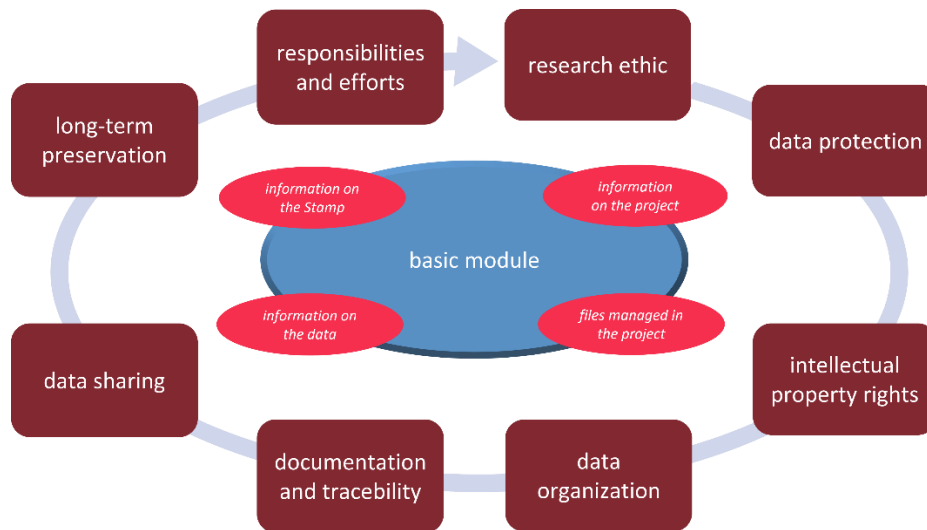


Figure 1: The Modules of the Stamp

To support reaching each minimal condition, content modules include *checklists*, outlining how to manage data appropriately. Within these checklists three types of auxiliary materials are referenced. Legal advice on data protection and intellectual property rights provides a deeper insight in such regulations. Guidelines refer to external guidance, best practise advice and templates of research associations, funders and repositories in educational research, the social sciences and beyond. Use cases exemplify projects in educational research, their various challenges in relation to RDM, and how these challenges were overcome.

In comparison to traditional data management plan templates, the Stamp comprises several advantages. First, instead of asking questions on various aspects of data management, the Stamp provides answers in terms of its checklists. It illustrates one way through a project's RDM, serving as a planning and a reporting tool on RDM that can be used, e.g., in funding applications or project reports. Second, the Stamp provides discipline-specific, tailored guidance in terms of its auxiliary materials, supporting RDM in educational research. Finally, the Stamp is designed to ensure shareable data, according to the idea of Open Science and the FAIR Data Principles, following up on requirements from, e.g., funding agencies, on sustainable research data.

The Usability of the Stamp Outside Educational Research

Although the Stamp provides tailored, discipline-specific guidance, it can be used outside educational research to a great extent. According to the expectation that RDM is primarily a matter of data, methods, and content, we organized two workshops with representatives of other social sciences disciplines as well as from research disciplines beyond the social sciences. Funded by KonsortSWD (2023), the workshops brought together a wide range of RDM experts from different disciplines, dealing with wide variety of data and discipline-specific requirements on RDM and data sharing.

In sum, participants agreed on the universal character of minimal conditions, reflecting requirements of good scientific practise and replicable research. Some minimal conditions might be far reaching for some disciplines, e.g., data protection regulations are only of relevance when processing personal data. Other minimal conditions might be short reaching. For

example, the Stamp does not cover patent rights, which might be of interest, e.g., in engineering sciences.

In addition, there was consensus on the usability of checklists, at least to some degree. Educational research is multidisciplinary, characterized by a large variety of data and a highly sensitive research population (Meyermann et al., 2017). Both characteristics foster the applicability of checklists, outside educational research. For example, checklists on data protection can be used, whenever processing personal data, irrespectively of the discipline. Likewise, checklists on, e.g., documenting distinct types of data can be employed, when documenting the same type of data, such as a data matrix, processed with similar methods.

Conclusions

In origin, the Stamp was designed for educational research, as obvious in its auxiliary materials and the terminology used, which is in line with the educational research community. When adapting the Stamp to other disciplines, minimal conditions must be adapted, first, to fit additional requirements of the respective discipline. Second, the terminology used in the Stamp needs to be 'translated', according to the language and terminology used in the respective discipline or community. Third, checklists must be adapted, e.g., by adding further types of data to be managed or further aspects of RDM, such as dealing with patent rights. Finally, auxiliary materials need to be replaced, according to guidance of the respective discipline.

Adapting the Stamp to other disciplines improves our understanding of RDM across disciplines. It highlights similarities and makes differences visible, fostering harmonization of RDM. To further elaborate the usability of the Stamp outside educational research and to examine how it supports developments towards a cross-disciplinary RDM, we started a second short-term project in 2023. Funded by KonsortSWD (2023), we will examine the usage of the Stamp in academic consulting and adapt it to institutional setting (Künstler-Sment 2023).

Competing interests

The authors declare that they have no competing interests.

Funding

The project DDP-Bildung was funded by the German Federal Ministry of Education and Research (grant number: 16QK01).

References

1. DDP-Bildung, and German Network of Educational Research Data (VerbundFDB). 2023. Stamp nutzen – Standardisierter Datenmanagementplan für die Bildungsforschung. Version 0.9. Frankfurt am Main: DIPF | Leibniz Institute for Research and Information in Education. Available at <https://www.forschungsdaten-bildung.de/stamp-nutzen>, last access 25 April 2023.
2. Grootveld, M., et al. 2018. OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans. Version 1.0.0. <http://doi.org/10.5281/zenodo.1120245>.
3. KonsortSWD. 2023 Consortium for the Social, Behavioural, Educational and Economic Sciences in the National Research Data Infrastructure (NFDI). Available at <https://www.konsortswd.de/>, last access 25 April 2023.
4. Künstler-Sment, J. 2023. Der Stamp in der Beratung. Available at https://www.iqb.hu-berlin.de/fdz/Projekte/Flyer_Stamp.pdf, last access 25 April 2023.

5. Meyermann, A., et al. 2017.: Der Verbund Forschungsdaten Bildung – Eine Forschungsdateninfrastruktur für die empirische Bildungsforschung. Available at https://www.konsortswd.de/wp-content/uploads/RatSWD_WP_266.pdf, last access 25 April 2023.
6. Netscher, S, I. Anders, and Ch. Henzen. 2022. Activities and Challenges in Developing Discipline-Specific Data Management Plan Templates: From Vertical to Horizontal Integration of RDM Practices. *Bausteine Forschungsdatenmanagement*, Nr. 1 (March 2022). pp. 13-25. <https://doi.org/10.17192/bfdm.2022.1.8371>.
7. Science Europe. 2018. Science Europe Guidance Document Presenting a Framework for Discipline-Specific Research Data Management. <https://doi.org/10.5281/zenodo.4925906>.
8. Wilkinson, M. D., et al. 2016. The FAIR Guiding Principles for Management and Stewardship. *Scientific Data* 3. <https://doi.org/10.1038/sdata.2016.18>.

Linked Open Research Data for Social Science

A concept registry for granular data documentation

Pascal Siegers¹[\[https://orcid.org/0000-0001-7899-6045\]](https://orcid.org/0000-0001-7899-6045), Antonia May¹[\[https://orcid.org/0000-0002-4979-0977\]](https://orcid.org/0000-0002-4979-0977), Claudia Saalbach²[\[https://orcid.org/0000-0002-1748-908X\]](https://orcid.org/0000-0002-1748-908X), Jana Nebelin², Dagmar Kern¹[\[https://orcid.org/0000-0003-1794-625X\]](https://orcid.org/0000-0003-1794-625X), Andreas Daniel³[\[https://orcid.org/0000-0002-0111-8858\]](https://orcid.org/0000-0002-0111-8858), Ben Zapilko¹[\[https://orcid.org/0000-0001-9495-040X\]](https://orcid.org/0000-0001-9495-040X), Fakhri Momeni¹[\[https://orcid.org/0000-0002-5572-575X\]](https://orcid.org/0000-0002-5572-575X), Knut Wenzig²[\[https://orcid.org/0000-0002-2259-0203\]](https://orcid.org/0000-0002-2259-0203), and Jan Gobel²[\[https://orcid.org/0000-0002-3243-1935\]](https://orcid.org/0000-0002-3243-1935)

¹ GESIS Leibniz-Institute for the Social Sciences, Germany

² Deutsches Institut für Wirtschaftsforschung eV

³ German Centre for Higher Education Research and Science Studies

Abstract. The re-use of research data is an integral part of research practice in the social and economic sciences. To find relevant data, researchers need adequate search facilities. However, a comprehensive, thematic search for research data is made more difficult by inconsistent or missing semantic indexing of data at the level of social science concepts (e.g., representing the theory language). Either the data is not documented at a granular level, or primary investigators use their ad-hoc terminology to describe their data. Consequently, researchers have to make great efforts to find relevant or comparable data. From the user's perspective, the lack of theory language in data documentation impedes effective data searches and thus significantly limits the research potential of existing data collections. Because there is currently no semantic model for indexing the data content, the specific challenge for improving data search lies in establishing concept-based indexing of research data. Research infrastructures need technology for the harmonized semantic indexing of their research data. The LORD concept registry aims at closing this gap by developing a registry of sociological and economic concepts and, following the FAIR principles, making this concept registry generally available to the scientific community. As a first step, we developed a basic data model for the Concept Registry using United Modeling Language (UML). All links between are created and managed in the form of so-called RDF triples. An annotation application allows for linking specific questions/variables to concepts. The application also includes the two SKOS-compliant thesauri, "Thesaurus Social Sciences" (TheSoz) and "Standard Thesaurus Economics" (STW) but could be extended to other resources like ELSST.

We illustrate the application of the LORD concept registry with examples from three large-scale survey programmes (German Socio-Economic Panel, German General Social Survey, National Academics Panel Study). The initial focus is on variables and questions with overlapping content in the three survey programmes, as they form a sound basis for cross-linking with concepts.

Keywords: Linked Open Data, Indexing, Interoperability, Survey research

Data availability statement

Data produced for this study is not ready for publication yet. And there is no practical use of it.

Funding

This presentation received a funding grant from the German Science Foundation (Grant Number 464413245).

1st Conference on Research Data Infrastructure

Humanities and Social Sciences

<https://doi.org/10.52825/CoRDI.v1i.393>

© Authors. This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

Published: 07 Sept. 2023

The Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI)

Better Infrastructure, Better Science, Better Society

Tom Emery^{1,2}[<https://orcid.org/0000-0001-6137-9577>], Kasia Karpinska^{1,2}[<https://orcid.org/0000-0003-3208-8046>],
Angelica Maineri^{1,2}[<https://orcid.org/0000-0002-6978-5278>], and Lucas van der Meer^{1,2}[<https://orcid.org/0000-0003-4415-678X>]

¹ Erasmus University Rotterdam – Erasmus School of Social and Behavioral Sciences

² Open Data Infrastructure for the Social Science and Economic Innovations (ODISSEI)

Abstract. The Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI) equips social scientists in the Netherlands with the data, tools, and skills that are necessary to answer groundbreaking questions for scientific and policy making purposes. With a variety of use cases to pick from, we aim at engaging in a discussion with other Research data infrastructures to identify synergies but also challenges ahead.

Keywords: Social Sciences; FAIR; data infrastructure.

1. Extended abstract

The Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI) equips social scientists in the Netherlands with the data, tools, and skills that are necessary to answer groundbreaking questions for scientific and policy making purposes. ODISSEI aspires to create a federated data infrastructure which seamlessly connects researchers with the data and facilities provided by member organisations. The experience gained by ODISSEI in the past six years contributes to a wider discussion on the role of Research Data Infrastructure in the Social Sciences. ODISSEI's programme of work is subdivided into four work streams (see Figure 1) which represent four separate ways in which ODISSEI serves the research community [1].

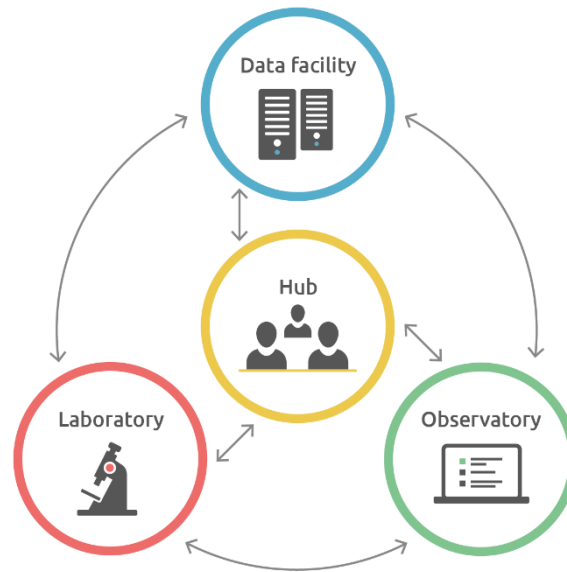


Figure 1. Conceptual representation of ODISSEI Infrastructure

First, the Data Facility ensures that researchers can find, access, and link the data that they need. At the centre of this work is the ODISSEI Secure Supercomputer which allows researchers to analyse complex and rich data from Statistics Netherlands and other ODISSEI data providers in a secure, yet computationally powerful environment (see examples of research conducted on the OSSC in [2] and [3]). ODISSEI also provides grants and support to access this data to facilitate new research. Moreover, the ODISSEI Portal makes diverse data sources available through a unique search interface by leveraging metadata [4]. ODISSEI is also leading the development of a generic Secure Analytical Environment (SANE) which makes it possible for data providers to securely share data with researchers whilst retaining overall control of data [5].

The Observatory supports and maintains key long-standing data collection efforts and participation in the international social science data collections. These include such studies as the European Social Survey, the Survey of Health, Ageing and Retirement in Europe and the Generations and Gender Programme. It also covers the Dutch Election Study which has been collecting data since 1971. This work stream is focused on providing a consistent, stable, and reliable stream of data for social scientists that could then be utilised across the infrastructure.

The Laboratory is where researchers can conduct their own experiments, primarily through the LISS panel, operated by Centerdata. ODISSEI provides financing for the core LISS panel but also provides access to researchers from ODISSEI member organisations to field their own questions to the LISS panel's representative and high-quality sample of over 4,500 households (see e.g. [6] and [7]). The Laboratory is also where future ODISSEI upgrades and enhancements are developed and prototyped.

Finally, the Hub is where researchers are provided with support, expertise, and guidance in the use of ODISSEI services and facilities. It includes an educational programme, community events, remote access grants, data stewardship, as well as a Social Data Science (SoDa) Team at Utrecht University who can provide high quality and intensive support to researchers looking to deploy computational and data science methods within ODISSEI.

As a community, ODISSEI not only delivers services but also develops and promotes standards and best practices for social science research. ODISSEI not only promotes the FAIR principles through the delivery of new search and access services, but also by requiring

adherence to FAIR from ODISSEI users. ODISSEI requires its users to act in accordance with the principles of responsible data science: Fair, Accurate, Confidential and Transparent (FACT). ODISSEI also supports open science by facilitating inclusion, sharing, and equity through its work [8].

ODISSEI also established, develops, and maintains a FAIR Expertise Hub to support communities of data providers in improving their FAIRness. An important instrument for the FAIR Expertise Hub is the FAIR Implementation Profile (FIP), a collection of decisions and plans made by a community about how to achieve FAIRness [9, 10]. A FIP comes with an easy to use wizard and accompanying workshop provided by GO-FAIR. The hub helps data communities in (1) establishing their plans, (2) to agree on their FAIR-enabling resources, and to (3) achieve a substantial increase of FAIRness. The project partners will (4) create alignment with international standards and (5) between communities. Explicit FAIR Implementation Profiles (6) facilitate software developers.

Looking ahead, ODISSEI will collaborate with CLARIAH (the Dutch infrastructure for digital humanities) in the SSHOC-NL project (the full proposal is available online, see [11]). SSHOC-NL aspires to develop an SSH-wide digital infrastructure that is interoperable, allowing its data, tools and services to be shared, linked, and combined in imaginative and ground-breaking ways.

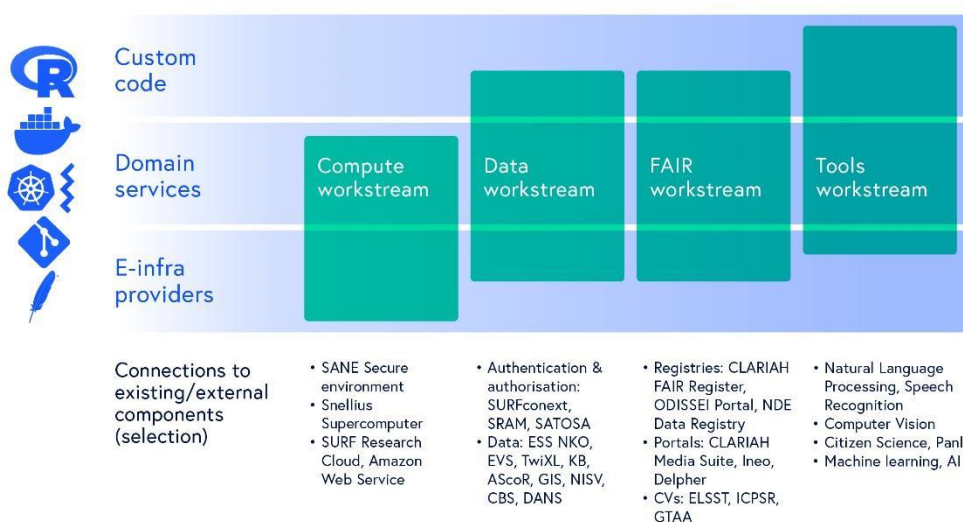


Figure 2. SSHOC-NL Architecture (see [11]).

With a variety of use cases to pick from, we aim at engaging in a discussion with other Research data infrastructures to identify synergies but also challenges ahead.

Data availability statement

The submission is not based on data.

Competing interests

The authors declare that they have no competing interests.

Funding

ODISSEI received a grant by the Dutch Research Council (NWO) under grant number 184.035.014.

Acknowledgement

We are grateful to all the partners in the ODISSEI project and to the members of the Management, Supervisory and Advisory Board.

References

1. T. Emery, P. Dykstra, and L. van der Meer, 'ODISSEI Annual Report 2021-2022', Zenodo, Dec. 2022. doi: <https://doi.org/10.5281/zenodo.7413577>.
2. E. L. de Zeeuw et al., 'Safe Linkage of Cohort and Population-Based Register Data in a Genomewide Association Study on Health Care Expenditure', *Twin Research and Human Genetics*, vol. 24, no. 2, pp. 103–109, 2021, doi: <https://doi.org/10.1017/thg.2021.18>.
3. A. Petrović, 'Multiscale spatial contexts and neighbourhood effects', Delft University of Technology, 2020. Accessed: Nov. 25, 2021. [Online]. Available: <https://journals.open.tudelft.nl/abe/article/view/5194>
4. T. Emery, R. Braukmann, M. Wittenberg, J. van Ossenbruggen, R. Siebes, and L. van de Meer, 'The ODISSEI Portal: Linking Survey and Administrative Data', Dec. 2020. doi: <https://doi.org/10.5281/zenodo.4302096>.
5. L. van der Meer, 'SANE - Secure ANalysis Environment @PDI-SSH', Jan. 26, 2023. doi: <https://doi.org/10.5281/zenodo.7562026>.
6. M. A. Yerkes, C. Remery, S. André, M. Salin, M. Hakovirta, and M. van Gerven, 'Unequal but balanced: Highly educated mothers' perceptions of work–life balance during the COVID-19 lockdown in Finland and the Netherlands', *Journal of European Social Policy*, p. 09589287221080411, Mar. 2022, doi: <https://doi.org/10.1177/09589287221080411>.
7. D. Yakar, Y. P. Ongena, T. C. Kwee, and M. Haan, 'Do People Favor Artificial Intelligence Over Physicians? A Survey Among the General Population and Their View on Artificial Intelligence in Medicine', *Value in Health*, Oct. 2021, doi: <https://doi.org/10.1016/j.jval.2021.09.004>.
8. User policy
9. E. Schultes, B. Magagna, K. M. Hettne, R. Pergl, M. Suchánek, and T. Kuhn, 'Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence', in *Advances in Conceptual Modeling*, G. Grossmann and S. Ram, Eds., in *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2020, pp. 138–147. doi: https://doi.org/10.1007/978-3-030-65847-2_13.
10. A. M. Maineri and S. Wang, 'FAIR yes, but how? FAIR Implementation Profiles in the Social Sciences', Dec. 2022, doi: <https://doi.org/10.5281/zenodo.74284>
11. P. Dykstra et al., 'Social Science and Humanities Open Cloud for the Netherlands (SSHOC-NL)', Feb. 2023, doi: <https://doi.org/10.5281/zenodo.7645356>.

Data publication for personalised health data

A new publication standard introduced by NFDI4Health

Juliane Fluck^{1,2} [<https://orcid.org/0000-0003-1379-7023>], Martin Golebiewski³ [<https://orcid.org/0000-0002-8683-7084>], Johannes Darms¹ [<https://orcid.org/0000-0001-5809-2276>] on behalf of the NFDI4Health consortium

¹ ZB MED Information Centre of Life Sciences, Germany

² University of Bonn, Germany

³ Heidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, Germany

Abstract. Health data collected in clinical trials and epidemiological as well as public health studies cannot be freely published, but are valuable datasets whose subsequent use is of high importance for health research. The National Research Data Infrastructure for Personal Health Data (NFDI4Health) aims to promote the publication of such health data without compromising privacy. Based on existing international standards, NFDI4Health has established a generic information model for the description and preservation of high-level metadata describing health-related studies, covering both clinical and epidemiological studies. As an infrastructure for publishing such preservation metadata as well as more detailed representation information of study data (e.g. questionnaires and data dictionaries), NFDI4Health has developed the German Central Health Study Hub. Content is either harvested from existing distributed sources or entered directly via a user interface. This metadata makes health studies more discoverable, and researchers can use the published metadata to evaluate the content of data collections, learn about access conditions and how and where to request data access. The goal of NFDI4Health is to establish interoperable and internationally accepted standards and processes for the publication of health data sets to make health data FAIR.

Keywords: Data publication, NFDI, FAIR, Search portal, epidemiological studies, clinical trials, personal health data

1. Background

Personal health data cannot be shared publicly due to data protection regulations. Even with the explicit consent of the study subjects for the subsequent use of the data for research purposes, the data collectors must exercise special care to protect and prevent misuse of the data. Unfortunately, to date, this has resulted in data being hidden within institutional boundaries and very limited data discoverability. Data reuse depends on whether data analysts are already familiar with the data collections or whether accompanying literature publications draw attention to the data sets. In the latter case, further information must often be obtained through direct contact with the data collectors. Undoubtedly, reuse of personal health data is a necessary step for further scientific advances. The COVID-19 pandemic has shown that discoverability of health data is enormously important and that information about health data should be published according to the FAIR principles.

2. Newly established publication standard by NFDI4Health

NFDI4Health [1] and the NFDI4Health Taskforce COVID-19 [2], in collaboration with various stakeholders, have created an information model that maps to domain-specific, as well as overarching metadata standards, and implemented this in a tailored technical infrastructure to enable data publication from personal health studies. Furthermore, the necessary materials and documentation were created to disseminate and engage the community in this effort.

2.1 Generic metadata model

A generic metadata standard tailored to the publication of clinical, epidemiological, and public health studies has been developed [3] (see Fig. 1 for an overview). This metadata schema is modular: in addition to a common core module, it includes modules specific to certain sub-domains. These are currently focused on the NFDI4Health use cases and have been developed based on input from domain experts. The metadata combines common elements and their controlled vocabularies (value sets) and can be used to describe studies with their corresponding resources (e.g. instruments, data collections, documents, data dictionaries, etc.), as well as study design and access conditions. Even without direct access to the data, these metadata provide all the information needed to be considered FAIR [4]. This is similar to a paper publication where access to full text or the underlying data may only be available upon payment of a license fee. It therefore can be considered as data publication for which we can also assign a Digital Object Identifier (DOI).

Mappings to domain-overarching (e.g. DataCite [5]) and to health-domain specific standards, and other metadata schemas (e.g. the ECRIN Clinical Research Metadata Repository [6]), as well as clinical trial registries, such as the International Clinical Trials Registry Platform (ICTRP [7]) of the WHO, the German Clinical Trials Register (DRKS [8]), and ClinicalTrials.gov [9], allow interfacing to external resources. To further enhance the interoperability, we also make use of health-domain specific ontologies (e.g. SNOMED CT [10]) and data interoperability standards, e.g. by implementing the metadata schema as profile in HL7 FHIR [11].

2.2 The German Central Health Study Hub

The German Central Health Study Hub (Health Study Hub for short) was initially developed to improve the findability of German COVID-19 studies (see Fig. 2). From the start, the Health Study Hub offered not only searches of study- and document-related preservation metadata, but also advanced searching, filtering, and comparison of individual questionnaires or variables.

The Health Study Hub allows the capture of preservation information about a study and associated documents directly via a user-friendly web-based data capture template or an application programming interface. We are in the process of building further dedicated and interoperable interfaces to existing platforms or services of data holding organisations to be able to transfer and reuse metadata. Furthermore, the Health Study Hub offers the possibility to publish study documents individually, especially the publication of questionnaires and variable catalogs is desired as those contain the most detailed information about available data.

The Health Study Hub contains over 1600 data assets (1522 studies and 107 study documents) with the majority (1578) related to Covid-19. The scope of the system is currently being expanded to include all German clinical and epidemiological studies. In addition to the large number of automatically integrated resources, more than 300 items were entered directly by experts and were previously unavailable. We expect these numbers to increase significantly as a result of various projects already initiated within the NFDI4Health project.

The technical implementation of the Health Study Hub was made possible by reusing and integrating existing software systems, in particular the Maelstrom Research Group's Mica software [12] and the Dataverse software [13] of the Harvard Institute for Quantitative Social Science.

Additionally, NFDI4Health is working closely with the Maelstrom Research group, which provides the majority of studies and study catalogs worldwide, to further establish this type of data publication internationally.

2.3 NFDI4Health support for data publication

Besides publication guidelines, that explain the process and concepts of publication of personal health data, NFDI4Health provides further information and training related materials on the subject. Additionally, workshops and hands-on trainings tailored to subdomains (e.g. nutritional epidemiologic) were conducted and further are planned. Ultimately, data stewards provide on-demand detailed support related to data publication.

3. Conclusions and Future Work

Building blocks to enable the publication of personal health data, a tailored and interoperable metadata model, the adaptation of standards and the technical implementation in the German Central Health Study Hub are available to the community. In addition, documentation, training and information on the publication process have been created and shared. Acceptance and use of the publication process is key to success and so close interaction with the community is crucial. The NFDI4Health approach points the way to FAIR publication of health data without disclosing protected data. In order to establish the publication of health data as a national and international standard, we are in a process of exchange with different national and international stakeholders and invite all interested parties to test and evaluate and provide feedback on the products created. We are keen to adapt and enrich the information model and services to integrate further needs. To this end, further mappings of the metadata model to other domain-specific metadata standards are planned, such as those of the German Human Genome-Phenome Archive (GHGA), the EU Clinical Trials Portal CTIS or the European Rare Disease Registry Infrastructure (ERDRI). Mappings to common metadata schemas such as the draft specification of the Data Documentation Initiative Cross-Domain Integration (DDI-CDI) or the J-PAL gold standard from MIT will also be considered, leading to further interoperability.

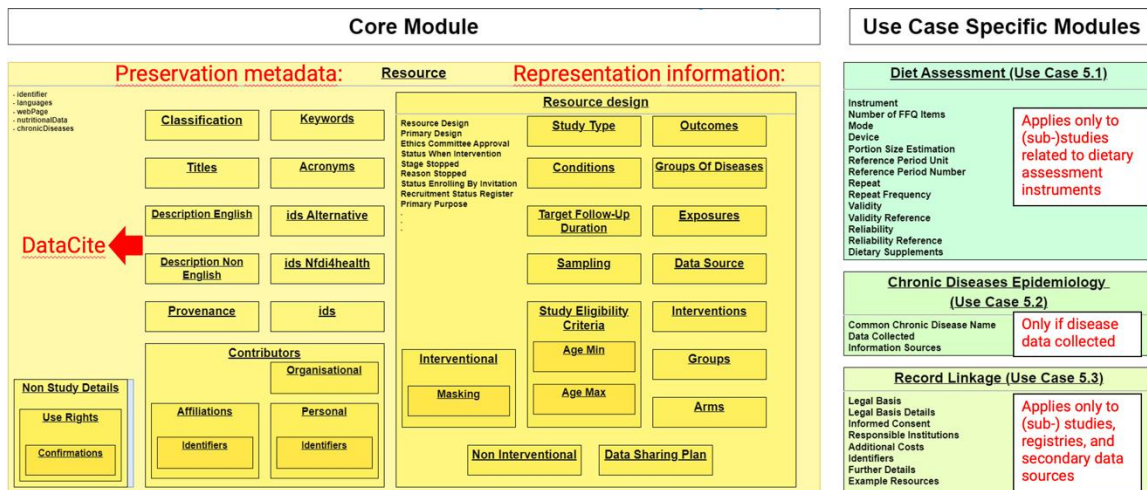


Figure 1. Subdomain-overarching NFDI4Health metadata schema designed in a modular way (schematic sketch)

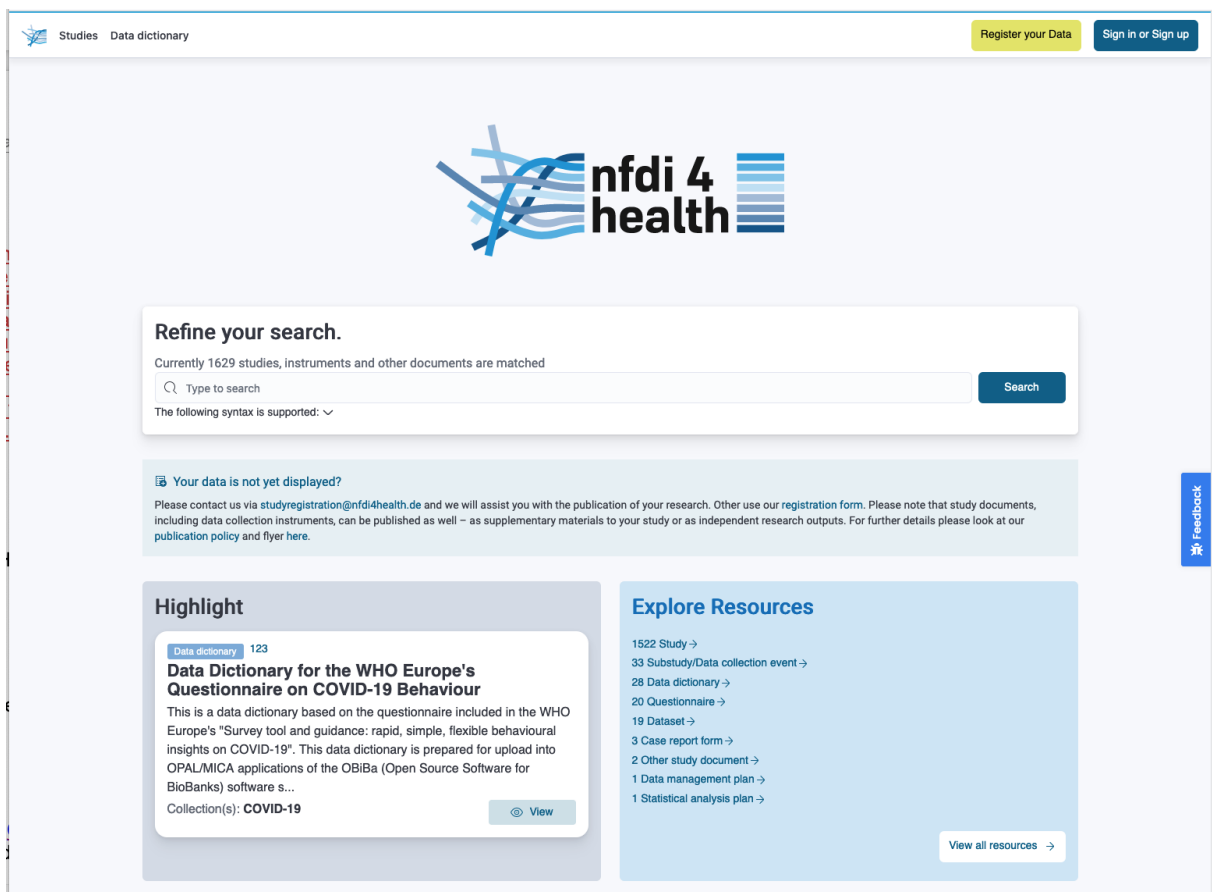


Figure 2. Screenshot of German Central Health Study Hub

Data availability statement

All data can be found in the German Central Health Study Hub or is referenced and available as Open-Source Publication.

Author contributions

JF, JD, MG prepared, reviewed and finalised the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

We acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG) through projects no. 442326535 (NFDI4health), and 451265285 (NFDI4health TF COVID19), as well as the Klaus Tschira Foundation (KTS).

Acknowledgement

We thank the whole NFDI4Health consortium for their valuable input.

References

1. Fluck, J., Lindstädt, B., Ahrens, W., Beyan, O., Buchner, B., Darms, J., Depping, R., Dierkes, J., Neuhausen, H., Müller, W., Zeeb, H., Golebiewski, M., Löffler, M., Löbe, M., Meineke, F., Klammt, S., Fröhlich, H., Hahn, H., Schulze, M., Pischon, T., Nöthlings, U., Sax, U., Kusch, H., Grabenhenrich, L., Schmidt, C.O., Waltemath, D., Semler, S., Gehrke, J., Kirsten, T., Praßer, F., Thun, S., Wieler, L., Pigeot, I., "NFDI4Health – Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten". BFDm, Nr. 2, S. 72–85 (2021). DOI: <https://doi.org/10.17192/bfdm.2021.2.8331>
2. Schmidt CO, Fluck J, Golebiewski M, Grabenhenrich L, Hahn H, Kirsten T, u. a. COVID-19-Forschungsdaten leichter zugänglich machen – Aufbau einer bundesweiten Informationsinfrastruktur, Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2021;64(9):1084–92 (2021). DOI: <https://doi.org/10.1007/s00103-021-03386-x>
3. Abaza H, Klopfenstein SAI, Golebiewski M, Schmidt CO, Shutsko A, Vorisek CN, Darms J., NFDI4Health Task Force COVID-19, NFDI4Health. "Metadata schema of the NFDI4Health and the NFDI4Health Task Force COVID-19 (V3_0)". (2022). DOI: <https://doi.org/10.4126/FRL01-006439110>.
4. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
5. DataCite. . Last access: April, 21 2023.
6. ECRIN Clinical Research Metadata Repository. <https://ecrin.org/clinical-research-metadata-repository> Last access: April, 21 2023.
7. "International Clinical Trials Registry Platform (ICTRP)" <https://www.who.int/clinical-trials-registry-platform>. Last access: April, 21 2023.
8. Deutsches Register Klinischer Studien (DRKS) - German Clinical Trials Register <https://drks.de/search/de>. Last access: April, 21 2023.
9. U.S. National Library of Medicine (NLM) at the U.S. National Institutes of Health (NIH), ClinicalTrials.gov. <https://clinicaltrials.gov>. Last access: April, 21 2023.
10. SNOMED International, "SNOMED CT" <https://www.snomed.org/>. Last access: April, 21 2023.
11. Health Level Seven International (HL7), "Fast Healthcare Interoperability Resources (FHIR)" <https://www.hl7.org/fhir/>. Last access: April, 21 2023.

12. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V., "Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination", *International Journal of Epidemiology*, 46(5):1372–8 (2017). DOI:<https://doi.org/10.1093/ije/dyx180>.
13. King, G., "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing", *Sociological Methods & Research*, 36(2), 173–199 (2017). DOI:<https://doi.org/10.1177/0049124107306660>.

The NFDI4Health – Task Force COVID-19

Iris Pigeot¹[\[https://orcid.org/0000-0001-7483-0726\]](https://orcid.org/0000-0001-7483-0726), Juliane Fluck^{2,3}[\[https://orcid.org/0000-0003-1379-7023\]](https://orcid.org/0000-0003-1379-7023), Johannes Darms²[\[https://orcid.org/0000-0001-5809-2276\]](https://orcid.org/0000-0001-5809-2276), and Carsten Oliver Schmidt⁴[\[https://orcid.org/0000-0001-5266-9396\]](https://orcid.org/0000-0001-5266-9396) on behalf of the NFDI4Health – Task Force COVID-19

¹ Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany

² Information Centre for Life Sciences, Köln, Germany

³ University of Bonn, Bonn, Germany

⁴ Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

Keywords: First keyword, Second, Third keyword

COVID-19 posed one of the greatest challenges to individuals and societies worldwide in recent decades. Public health research, epidemiological and clinical studies were essential to track the spread of SARS-CoV-2 responsible for the pandemic and its variants, to better understand the consequences for health and social life, and to identify effective treatment and vaccination methods. Such studies provided policy makers, industry, health care providers, and society with an empirical basis for containing and managing the pandemic and for making decisions that were based on the most recent data. Therefore, the COVID-19 pandemic excellently illustrates the relevance of data sharing and the importance of providing an effective infrastructure. From the researchers' perspective, there were significant challenges associated with this request. In a very short time, numerous projects, studies, and networks had emerged to investigate the pandemic, making it increasingly difficult to maintain an overview. Such an overview would have been essential to coordinate research activities, avoid unplanned duplication of research, and to implement studies in a harmonized manner.

Hurdles already varied vastly when trying to find ongoing studies. Due to the existing obligation to register clinical trials in registries, their well-structured metadata are available. In contrast, the situation for epidemiological and public health studies was much less clear. Although there were several national and international overviews on the Internet, e.g., a COVID-19 research registry of the American Society for Microbiology [1], the COVID-19 research overview of the Medical Informatics Initiative [2], or the German Data Forum [3], these overviews were inconsistent in scope, timeliness, and depth of information.

It became even more difficult when a detailed insight into protocols, survey instruments, item banks, and other study documents was requested across studies. Only in isolated cases did projects provide access to relevant information. For example, the German Corona Consensus Dataset (GECCO) is a positive example of harmonized data collection based on international medical IT standards using a coordinated core dataset. It was created in the Network University Medicine (NUM) [4], which primarily coordinates hospital-related research [5].

In addition, during the course of the pandemic, other obstacles became apparent that hampered efficient research: Although in some cases the same individuals were included in different studies and further health data of these individuals were stored, e.g., by health insurance companies, there are insufficient options to link these data on an individual level. This limits the possibilities of obtaining a sufficiently comprehensive picture of disease occurrence, on progression prognosis or vaccination consequences.

There was also a need to share even preliminary research results in order to respond adequately to the pandemic. As a result, the publication of preprints and the importance of preprint servers for sharing results without peer review increased, as the established peer-reviewed publication process was too slow during this period.

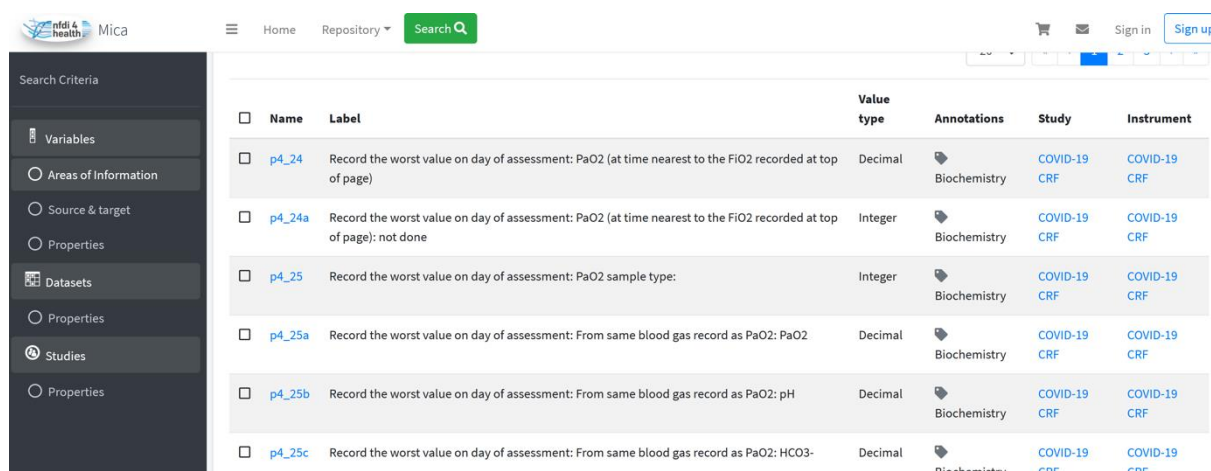
In summary, despite positive examples, the German clinical studies on COVID-19 and corresponding datasets in epidemiology and public health have only insufficiently met the requirements of the so-called FAIR principles [6]. To address this shortcoming, the NFDI4Health Task Force COVID-19 [7] was established as part of NFDI4Health [8]. In doing so, the NFDI4Health Task Force COVID-19 focused on research related to patients with COVID-19 as well as the public health consequences of the pandemic outbreak on the general population.

This project led to a range of key outcomes that meanwhile serve as a basis for NFDI4Health developments. For one part, a first version of a central search portal, the German Central Health Study Hub COVID-19, has been established. This application allows users to search, explore and retrieve study-related information. The main focus of the application is the consolidation of distributed information in order to reduce search time. It has since been expanded to include additional datasets and functionality to form the NFDI4Health German Central Health Study Hub (<https://csh.nfdi4health.de>) (Figure 1). The search engine is complemented by an instrument portal function. This allows COVID-19 related survey instruments to be searched semantically. The Maelstrom taxonomy [11],[12] is used for this purpose. This taxonomy consists of 18 domains, such as socio-demographic and economic characteristics and diseases, which in turn are divided into 135 subdomains (e.g. ICD domains). The following link (<https://mica.covid19.studyhub.nfdi4health.de/>) provides direct access to the function.

The screenshot shows the search interface of the German Central Health Study Hub. At the top, there are navigation links for 'Studies' and 'Data dictionary', along with buttons for 'Register your Data' and 'Sign In or Sign up'. The main heading reads 'Refine your search. Currently 1504 studies, instruments and other documents are matched'. Below this is a search input field with a 'Search' button and a 'Reset' button. A dropdown menu indicates 'The following syntax is supported:'. The left sidebar contains filters for 'Collection(s)', 'Type of the resource', and 'Specification of study type'. The main content area displays two study results, each with a 'View' button. The first study is 'My childhood - Your childhood Study on the Influence of Mothers' Childhood Experiences on their Children from Birth to School Age' (TRANS-GEN). The second study is 'Impact of the COVID-19 pandemic on surgical indication and postoperative course in patients with cholecystectomy (CHE)'. A 'Feedback' button is visible on the right side of the page.

Figure 1. German Central Health Study Hub – COVID-19 collection

An important underlying activity was the development of a metadata model to describe studies and related resources. Compatibility with existing registries and standards such as the ICTRP [9] and the German Register of Clinical Studies (DRKS) [10], the Minimum Information About Biobank Data Sharing (MIABIS), the Maelstrom data model and the DataCite metadata schema was taken care for. Mappings against HL7 FHIR and CDISC ODM standards were conducted. These results have also been incorporated into the NFDI4Health initiative and are already being expanded.



<input type="checkbox"/>	Name	Label	Value type	Annotations	Study	Instrument
<input type="checkbox"/>	p4_24	Record the worst value on day of assessment: PaO2 (at time nearest to the FiO2 recorded at top of page)	Decimal	Biochemistry	COVID-19 CRF	COVID-19 CRF
<input type="checkbox"/>	p4_24a	Record the worst value on day of assessment: PaO2 (at time nearest to the FiO2 recorded at top of page); not done	Integer	Biochemistry	COVID-19 CRF	COVID-19 CRF
<input type="checkbox"/>	p4_25	Record the worst value on day of assessment: PaO2 sample type:	Integer	Biochemistry	COVID-19 CRF	COVID-19 CRF
<input type="checkbox"/>	p4_25a	Record the worst value on day of assessment: From same blood gas record as PaO2: PaO2	Decimal	Biochemistry	COVID-19 CRF	COVID-19 CRF
<input type="checkbox"/>	p4_25b	Record the worst value on day of assessment: From same blood gas record as PaO2: pH	Decimal	Biochemistry	COVID-19 CRF	COVID-19 CRF
<input type="checkbox"/>	p4_25c	Record the worst value on day of assessment: From same blood gas record as PaO2: HCO3-	Decimal	Biochemistry	COVID-19 CRF	COVID-19 CRF

Figure 2. Instrument portal

The initiative also set up and elaborated a range of other services such as tools for assessing imaging quality, tools for assessing data quality, and a semantic search engine for preprints [13]. Regarding the latter, metadata were queried from the preprint servers medRxiv, bioRxiv, ChemRxiv, ResearchSquare, arXiv and Preprints.org and transferred to a shared data schema. A terminology was created to identify viral SARS-CoV-2 proteins using a dictionary-based algorithm, a web-based user interface, and a programming interface were developed to provide users with semantic search functionalities.

This infrastructure makes it easier to find research and its results on SARS-CoV-2 and COVID-19 from public health, epidemiology and clinical studies. The described developments, which were initiated as part of the NFDI4Health Task Force COVID-19 [12] are also relevant beyond COVID-19, as the challenges addressed are generic for finding and exploiting research data. Thus, the Task Force COVID-19 may be regarded as a kind of microcosm that served as a blueprint for the activities required to successfully implement the services and standards developed by the NFDI4Health.

Author contributions

All authors collaborate in the NFDI4health project. IP, COS prepared the manuscript, all authors reviewed and finalized.

Competing interests

The authors declare that they have no competing interests.

Funding

We greatly appreciate the funding from the Deutsche Forschungsgemeinschaft (DFG) through project no. 451265285 (NFDI4health - Task Force COVID-19)

References

1. American Society for Microbiology (2022), COVID-19 research registry. <https://asm.org/COVID/COVID-19-Research-Registry/Epidemiology>. Last access: April 15, 2023
2. Medical Informatics Initiative (2021), Overview of Covid-19 research. <https://www.medizininformatik-initiative.de/en/node/410>. Last access: April 15, 2023

3. German Data Forum (2023), Studien zur Corona-Pandemie. <https://www.kon-sortswd.de/ratswd/themen/corona/studien/>. Last access: April 15, 2023
4. Network University Medicine (2023), Homepage. <https://www.netzwerk-universitaetsmedizin.de/>. Last access: April 15, 2023
5. Sass, J., Bartschke, A., Lehne, M. et al. (2020). The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. *BMC Med Inform Decis Mak* 20, 341. DOI: <https://doi.org/10.1186/s12911-020-01374-w>
6. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>
7. NFDI4Health (2023), Task Force COVID-19. <https://www.nfdi4health.de/en/task-force-covid-19.html>. Last access: April 15, 2023
8. NFDI4Health (2023), NFDI4Health – Eine Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten. <https://www.nfdi4health.de/en/>. Last access: April 15, 2023
9. World Health Organization (2023), ICTRP search portal. <https://www.who.int/clinical-trials-registry-platform/the-ictrp-search-portal>. Last access: April 15, 2023
10. German Clinical Trials Register (DRKS) (2023), Deutsches Register Klinischer Studien. <https://www.drks.de/>. Last access: April 15, 2023
11. Schmidt, C. O., Fluck, J., Golebiewski, M., et al. (2021). COVID-19-Forschungsdaten leichter zugänglich machen – Aufbau einer bundesweiten Informationsinfrastruktur [Making COVID-19 research data more accessible-building a nationwide information infrastructure]. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, 64(9), 1084–1092. DOI: <https://doi.org/10.1007/s00103-021-03386-x>
12. Schmidt CO, Darms J, Shutsko A, et al. (2021). Facilitating Study and Item Level Browsing for Clinical and Epidemiological COVID-19 Studies. *Studies in Health Technology and Informatics*. 2021 May;281:794-798. DOI: 10.3233/shti210284. PMID: 34042687.
13. Langnickel, L., Darms, J., Heldt, K., et al. (2022). Continuous development of the semantic search engine preVIEW: from COVID-19 to long COVID. *Database*, Volume 2022, baac048. DOI: <https://doi.org/10.1093/database/baac048>

Connecting National and International Data Infrastructures in Biodiversity Research

The Case of NFDI4Biodiversity, a German Consortium for Biodiversity, Ecology and Environmental Data

Barbara Ebert¹[\[https://orcid.org/0000-0003-3328-6693\]](https://orcid.org/0000-0003-3328-6693), Judith Sophie Engel² [\[0000-0001-8665-6382\]](https://orcid.org/0000-0001-8665-6382), Ivaylo Kostadinov¹[\[https://orcid.org/0000-0003-4476-6764\]](https://orcid.org/0000-0003-4476-6764), Anton Güntsch³[\[https://orcid.org/0000-0002-4325-4030\]](https://orcid.org/0000-0002-4325-4030) and Frank Oliver Glöckner^{2,4} [\[https://orcid.org/0000-0001-8528-9023\]](https://orcid.org/0000-0001-8528-9023)

¹ GFBio - German Federation for Biological Data, Germany

² MARUM - Center for Marine Environmental Sciences, University of Bremen, Germany

³ Center for Biodiversity Informatics and Collection Data Integration, Botanic Garden and Botanical Museum Berlin, Germany

⁴ AWI - Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Germany

Abstract. NFDI4Biodiversity is a consortium within the German National Research Data Infrastructure (NFDI) dedicated to data and services for biodiversity research and ecology. In the domain, well-developed international networks exist, with quite mature tools and data standards. These are used and disseminated in the NFDI4Biodiversity project in order to mobilise and publish data collected by national stakeholders according to the FAIR Guiding Principles for scientific data management and stewardship. The consortium partners provide methods and tools for archiving, publishing, searching and analysing data that are suitable for everyday use and have been tried and tested in practice. The consortium also functions as a forum for technical and legal matters of data and data-related workflows. In this way, the consortium is providing added value to the community regarding access to modern technologies and a comprehensive stock of biodiversity and environmental data.

Keywords: Biodiversity, Ecology, Community-based services, Research Data, NFDI, Germany, NFDI4Biodiversity

1. Setting the Scene

NFDI4Biodiversity is a consortium within the German National Research Data Infrastructure (NFDI) dedicated to data and services for biodiversity research and ecology [1]. The 50 partner organisations represent typical stakeholders in the domain: Research groups with a variety of methods and data types from taxonomic and ecosystem research, natural history collections with digitisation projects for data and objects, and the nature conservation domain, where Citizen scientists and highly specialised expert organisations contribute large amounts of structured observation data. IT service centres and informatics research groups bring important technical competencies to the group.

Data on the biodiversity of animals, plants and microorganisms typically include species observations recorded in the form of tables, photos, videos, specimens and audio files. Specimens are also frequently collected and preserved for further analysis and description.

Increasingly important is genetic information on the observed species, i.e. sequence data and data from other -omics methodology (metabolomics, glycomics, and transcriptomics), which provide information on biological functions. Contextual data on environmental conditions in the species' habitat, on land use or colonisation are also relevant. Data streams from continuous monitoring of environmental parameters by sensors/satellites bring on new infrastructure requirements.

Well-developed international networks exist, with quite mature tools and data standards (see Table 1). These are used and disseminated in the NFDI4Biodiversity project in order to mobilise and publish data collected by national stakeholders according to the FAIR Guiding Principles for scientific data management and stewardship. NFDI4Biodiversity also builds on services provided by the German Federation for Biological Data (GFBio e.V.), the German Network on Bioinformatics Infrastructure - Deutsches Netzwerk für Bioinformatik Infrastruktur (de.NBI), and the eight German GBIF nodes.

Table 1. International data initiatives with standards and tools of practical relevance to the current work (in alphabetical order)

Name	Scope	Tools and standards relevant in current work
BioCASE - Biological Collection Access Service	Network of primary biodiversity repositories	BioCASE protocol BioCASE monitor software BioCASE provider software
CETAF - Consortium of European Taxonomic Facilities	Object-related and taxonomic research, biological and geological collections	Best practices - CETAF stable identifiers
ELIXIR - European Research Infrastructure for the Life Sciences	Molecular and -omics research	European Nucleotide Archive ENA RDM-Kit - Research Data Management toolkit Bioschemas Policy and specification
GBIF - Global Biodiversity Information Facility	Data, common standards, open-source tools	Scientific Data Collection Hosted portals Integrated Publishing Toolkit
International Barcode of Life	Data and DNA-based tools	Barcode of Life Data System
RDA - Research Data Alliance	Recommendations and standards	I-ADOPT Framework
TDWG - Biodiversity Information standards	Standards and guidelines for the recording and exchange of data about organisms	Access to Biological Collection Data (ABCD) Schema

GFBio services are structured along the data life cycle, from the preparation of a data management plan to data archiving and publishing [2]. A data submission service is provided, where researchers can submit heterogeneous data files from their projects for curation and archiving in matching data archives, including the European Nucleotide Archive ENA. Any submission or consultation generates a ticket that is handled by a professional helpdesk team with representatives from the associated data centres.

2. The NFDI4Biodiversity approach

Mobilising data is facilitated if partners get added value for their own day-to-day business, for example the digitisation of work processes. In a use case project with the Bavarian Forest National Park, BEXIS (Information System of the Biodiversity Exploratories) was established as a hosted database solution for scientific data management in the national park's projects, facilitating data publications in the future. In the use case for a Living Atlas of Germany, partners are provided with BioCASE installations to facilitate data publications from local systems to the GBIF and GFBio data portals. In 2022, flexible funds from the NFDI4Biodiversity consortium were used to create the development of an open web service interface to the new checklist infrastructure of the National Red List Centre.

In order to increase our outreach to scientists, we introduced a "Front Office/Back Office model" in the Helpdesk support: Initial consulting by local data stewards at research performing organisations is linked to the subject-specific consulting services of the NFDI4Biodiversity network (see Fig. 1). The model was prototypically implemented with the Research Initiative for Biodiversity Conservation (FEEdA) [3]. We see the model as an effective building block for the professional management of data from large research collaborations as well as the "long tail of science" at universities and institutes.

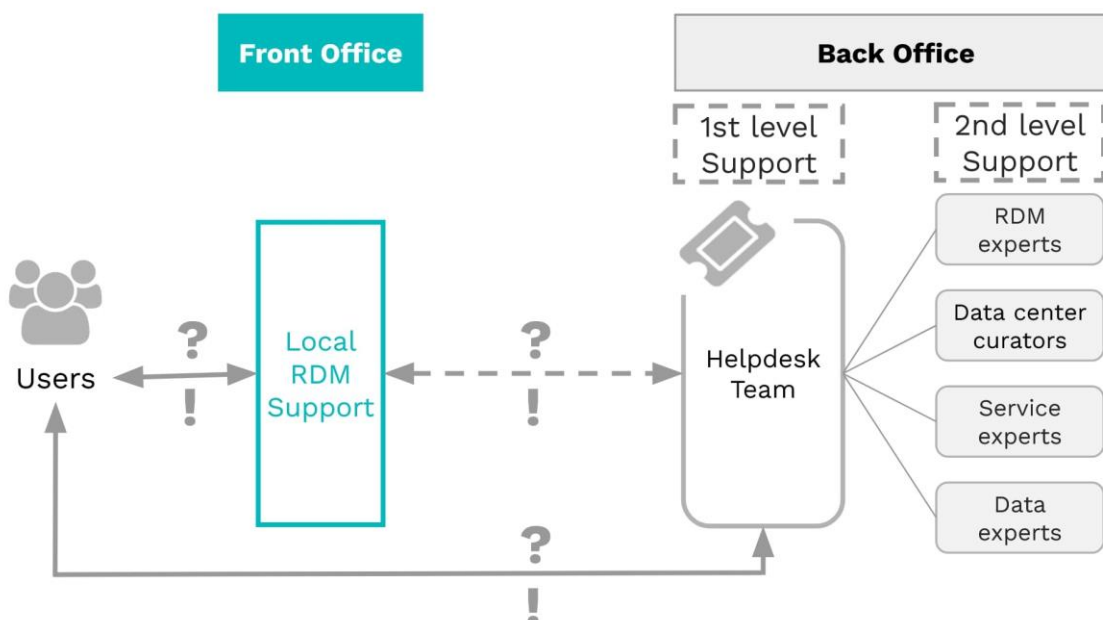


Figure 1. Front Office/Back Office support provided by the NFDI4Biodiversity Helpdesk (adapted from [3])

On a technical level, the vision of NFDI4Biodiversity is a federated, distributed IT infrastructure ("Commons") that provides users with access to tried-and-tested applications for data analysis and publication, in combination with tools for storage and semantic integration. Scientific IT centres in the network provide cloud infrastructure and storage and act as bridges into the network of high-performance computing centres.

3. Conclusions

By combining international tools and standards with a national service infrastructure, the NFDI4Biodiversity consortium is providing access to modern technologies and a comprehensive stock of biodiversity and environmental data. The consortium is also a forum for technical and legal matters of data and data-related workflows. With regard to future

development, we see two important lines of development: a) sustainable funding for the public core data resources in the life sciences; b) integration of data management structures in the national and international biodiversity monitoring programmes with the scientific research infrastructure landscape. In this way, we can mobilise and provide the extensive data that stakeholders from science, politics, economy and society need for better contributions to the conservation of global biodiversity.

Data availability statement

The submission is not based on data.

Underlying and related material

none

Competing interests

The authors declare that they have no competing interests.

Funding

NFDI4Biodiversity is funded by the German Research Foundation (DFG) within the framework of an agreement between the Federal Government and the Länder on the establishment and funding of the National Research Data Infrastructure (NFDI) of 26 November 2018 (grant number [442032008](#)).

Acknowledgement

We thank the dedicated staff and partners who help to shape and support the structures and results presented here. With regard to this article special thanks go to Merle Schwarten for editing the manuscript.

References

1. F. O. Glöckner et al. "NFDI4BioDiversity - A Consortium for the National Research Data Infrastructure (NFDI)", Zenodo, Jul 2020, <https://doi.org/10.5281/zenodo.3943645>
2. German Federation for Biological Data: GFBio Services, <https://www.gfbio.org/services/> (accessed on 2023-04-25)
3. FEdA is a 10-year funding program with 25-30 research projects and a central coordination office (www.feda.bio). For information on the NFDI collaboration see: J. Linares Gómez, B. Ebert, J. Eberhardt, K. Frohne, K. Sauerland, K. Mozygamba, B. Miller. "Collaborative Research Data Management Support for FEdA projects", Zenodo, Feb 2023, <https://doi.org/10.5281/zenodo.7624583>

The EOSC-Life WORKFLOW COLLABORATORY for the LIFE SCIENCES

Carole Goble¹[\[https://orcid.org/0000-0003-1219-2137\]](https://orcid.org/0000-0003-1219-2137), Finn Bacall¹[\[https://orcid.org/0000-0002-0048-3300\]](https://orcid.org/0000-0002-0048-3300), Stian Soiland-Reyes¹[\[https://orcid.org/0000-0001-9842-9718\]](https://orcid.org/0000-0001-9842-9718), Stuart Owen¹[\[https://orcid.org/0000-0003-2130-0865\]](https://orcid.org/0000-0003-2130-0865), Ignacio Eguinoa²[\[https://orcid.org/0000-0002-6190-122X\]](https://orcid.org/0000-0002-6190-122X), Bert Driesbeke²[\[https://orcid.org/0000-0003-0522-5674\]](https://orcid.org/0000-0003-0522-5674), Hervé Ménager³[\[https://orcid.org/0000-0002-7552-1009\]](https://orcid.org/0000-0002-7552-1009), Laura Rodriguez-Navas⁴[\[https://orcid.org/0000-0003-4929-1219\]](https://orcid.org/0000-0003-4929-1219), José M. Fernández⁴[\[https://orcid.org/0000-0002-4806-5140\]](https://orcid.org/0000-0002-4806-5140), Björn Grüning⁵[\[https://orcid.org/0000-0002-3079-6586\]](https://orcid.org/0000-0002-3079-6586), Simone Leo⁶[\[https://orcid.org/0000-0001-8271-5429\]](https://orcid.org/0000-0001-8271-5429), Luca Pireddu⁶[\[https://orcid.org/0000-0002-4663-5613\]](https://orcid.org/0000-0002-4663-5613), Michael R. Crusoe⁷[\[https://orcid.org/0000-0002-2961-9670\]](https://orcid.org/0000-0002-2961-9670), Johan Gustafsson⁸[\[https://orcid.org/0000-0002-2977-5032\]](https://orcid.org/0000-0002-2977-5032), Salvador Capella-Gutierrez⁴[\[https://orcid.org/0000-0002-0309-604X\]](https://orcid.org/0000-0002-0309-604X), and Frederik Coppens²[\[https://orcid.org/0000-0001-6565-5145\]](https://orcid.org/0000-0001-6565-5145)

¹ The University of Manchester, UK

² VIB-UGent Center for Plant Systems Biology, Belgium

³ Institut Pasteur, Paris

⁴ Barcelona Supercomputing Center, Spain

⁵ Albert-Ludwigs-University Freiburg, Germany

⁶ Center for Advanced Studies, Research and Development in Sardinia, Italy

⁷ Common Workflow Language & VU Amsterdam, Netherlands

⁸ Australian BioCommons, Australia

Abstract. Workflows have become a major tool for the processing of Research Data, for example, data collection and data cleaning pipelines, data analytics, and data update feeds populating public archives. The EOSC-Life Research Infrastructure Cluster project brought together Europe's Life Science Research Infrastructures to create an Open, Digital and Collaborative space for biological and medical research to develop a cloud-based Workflow Collaboratory. As adopting FAIR practices extends beyond data, the Workflow Collaboratory drives the implementation of FAIR computational workflows and tools. It fosters tool-focused collaborations and reuse via the sharing of data analysis workflows and offers an ecosystem of services for researchers and workflow specialists to find, use and reuse workflows. It's web-friendly Digital Object Metadata Framework, based on RO-Crate and Bioschemas, supports the description and exchange of workflows across the services.

Keywords: Fair Data; Computational Workflows; Digital Objects; Data Intensive Bio-Science; Fair Workflows; Fair Software

1. Background

Performing computational data processing using workflows has taken hold in the biosciences as the discipline becomes increasingly computational. Adopting FAIR practices extends beyond data to include workflows and the tools they use. The COVID-19 pandemic highlighted the importance of systematic and shared analysis of SARS-CoV-2 data processing and surveillance pipelines, data analytics at scale, and the reproducibility of computational processes [1].

The EOSC-Life Research Infrastructure Cluster project brought together Europe's Life Science Research Infrastructures to create an open, digital and collaborative space for biological and medical research. The Research Infrastructures, range from biobanking and clinical trials to plant phenotyping and bioimaging. A major development by EOSC-Life has been a cloud-based Workflow Collaboratory to drive the implementation of FAIR computational workflows [2] and foster tool-focused collaborations and reuse via the sharing of data analysis workflows.

1.1 The EOSC-Life Workflow Collaboratory Services

The Workflow Collaboratory offers an ecosystem of services for researchers and workflow specialists to find, use and reuse workflows, and deploy them using European Open Science Cloud (EOSC) infrastructure (Figure 1). The heterogeneity of the Research Infrastructures is reflected in the diversity of their data analysis practices and the variety of workflow management systems they use, including specialist platforms.

Workflow Managers and Execution systems include pre-existing and emerging workflow management systems. Currently 14 different workflow platforms are represented including Jupyter Notebooks [3] and Python scripts, general systems (e.g. Nextflow [4], Snakemake[5], Galaxy[6], CWL[7]) and specialist systems (e.g. SCIPION). Workflow execution platforms include Galaxy Europe and back-end services Sapporo and WfExS that execute workflows in different languages (e.g. CWL, Nextflow, snakemake) through a common interface. WfExS handles sensitive data securely.

Community Workflow Repositories both pre-existing and emerging, typically use Git, GitLab or GitHub. Curated collections include Galaxy's Intergalactic Workflow Commission and Nextflow's nf-core; others include project focused repositories.

Registries provide one stop to find and share containers (BioContainers [8]), tools (bio.tools [9]) and workflows (WorkflowHub [10]), and support FAIRness through rich metadata and inter-registration integration. WorkflowHub is system agnostic, with dedicated support for popular management systems. Galaxy and CWL workflows, and entries on WorkflowHub, are annotated with tool identifiers to link through to bio.tools entries, and bio.tools links to workflows that use a given tool.

Testing and Benchmarking services support the usability as well as reusability of tools and workflows as software [11]. LifeMonitor [10] monitors and triggers automated workflow tests and automated checks on metadata, and adherence to best practices on the workflow's source code Git repository. OpenEBench benchmarks tools, and monitors software quality as well as scientific benchmarking to help determine the precision, recall and other metrics of bioinformatics resources in unbiased scenarios.

Research Information System services plug into the wider services of the EOSC and scholarly communication to support publication, citation, and knowledge discovery. WorkflowHub's integration DataCite mints DOIs for workflows publication, and its integration with ORCID and citation.cff file format supports workflow author credit and citation. The WorkflowHub contributes to the DataCite PID Graph and OpenAIRE Research Graph.

Infrastructure services range from AAI (LS-Login¹) to cloud and cluster compute systems such as Galaxy's PULSAR network².

¹ <https://lifescience-ri.eu/ls-login/>

² <https://pulsar-network.readthedocs.io/en/latest/>

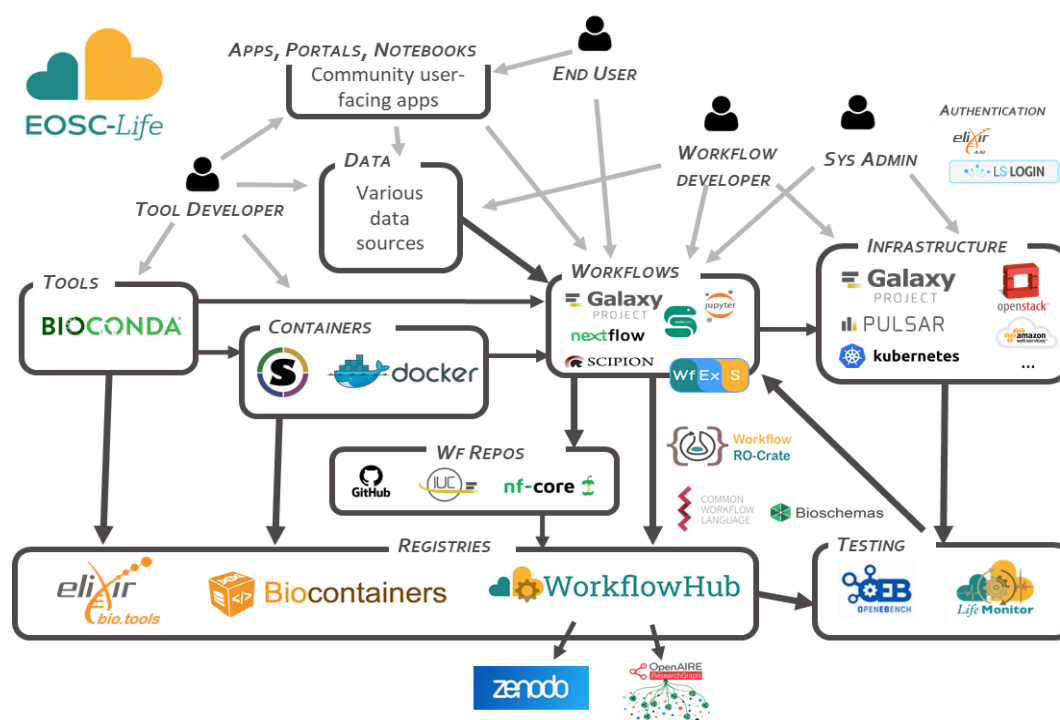


Figure 1. The EOSC-Life Workflow Collaboratory.

1.2 The Digital Object based Metadata Framework

A web-friendly metadata framework has been developed to support the description and exchange of workflows across the services.

Bioschemas [12] schema.org profiles for Computational Tool, Computational Workflow and Formal Parameter provide metadata about a workflow and its tools that are discipline independent, despite the “bio” prefix. The **EDAM Ontology** [13] adds informatics-specific metadata, such as strong typing of inputs and outputs and describes the overall workflow topics and operations to help find workflows.

The **Common Workflow Language** [7] is encouraged as a canonical workflow description to accompany the native workflow definitions when registering in the WorkflowHub. CWL represents the structure and steps of workflows in an interoperable way across workflow languages. Several workflow systems, such as Galaxy, support Abstract CWL for this purpose, though they are still executed using their native language on their native platforms.

RO-Crate [14], a community-developed standardised approach for FAIR Digital Objects [15], packages executable workflows, their components (e.g. example and test data), abstract CWL, diagrams and their metadata. RO-Crate makes workflows more readily re-usable, acting as the unit of currency of exchange between the services, recording the provenance of workflow runs and a format for archiving in public repositories such as Zenodo.

The GA4GH Tools Registry Service API supports the exchange of scientific tools and workflows and enables users to search for and retrieve metadata about registered tools, so that workflow execution platforms can search and import workflows from WorkflowHub and WorkflowHub can directly launch workflows on their platforms. Multi-language execution services such as Sapporo use the GA4GH Workflow Execution Service API.

2. Outlook

At the time of writing over 380 workflows have been registered in WorkflowHub from over 170 workflow teams. Adoption of the services has extended beyond Life Sciences to adoption by communities working in biodiversity, astronomy, astro-physics and climate change. These services continue to be supported by a new portfolio of Horizon Europe and national projects, and are sponsored by the European Life Science Research Infrastructures, ELIXIR, EuroBioImaging-ERIC, BBMRI-ERIC, and EU-IBISBA for their long-term sustainability. Many of the services are registered in the EOSC Service Catalogue and Marketplace³ and have been adopted outside Europe by the Australian BioCommons⁴.

Data availability statement

All workflows, tools and content are openly available. Software is open source. Standards are open. Available from the following: WorkflowHub: <https://workflowhub.eu/>; BioContainers: <https://biocontainers.pro/>; Bio.tools: <https://bio.tools/>; LifeMonitor: <https://www.lifemonitor.eu/>; OpenEBench: <https://openebench.bsc.es/>; Galaxy Europe: <https://usegalaxy.eu/>; Sapporo <https://github.com/sapporo-wes/sapporo>; WfExS: <https://github.com/inab/WfExS-backend>; Bioschemas: <https://bioschemas.org>; Common Workflow Language: <https://www.commonwl.org/>; EDAM Ontology: <https://edamontology.org/page>; RO-Crate: <https://www.researchobject.org/ro-crate/>; GA4GH Tools Research Service API: <https://ga4gh.github.io/tool-registry-service-schemas/>; GA4GH Workflow Execution Service API: <https://ga4gh.github.io/workflow-execution-service-schemas/docs/>

Underlying and related material

None

Author contributions

CG wrote the abstract and supervised the work. FB, SS-R, SO, IE, BD, HM, LR-N, JMF, BJ, SL, JG undertook the work. SC-G, FC supervised the work.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the European Union programmes Horizon 2020 under grant agreements H2020-INFRAEDI-02-2018 823830 (BioExcel-2), H2020-INFRAEOSC-2018-2 824087 (EOSC-Life), and Australian BioCommons which is enabled by NCRIS via Bioplatforms Australia funding.

³ <https://marketplace.eosc-portal.eu/>

⁴ <https://australianbiocommons.github.io/>

Acknowledgement

We acknowledge the WorkflowHub Club (<https://about.workflowhub.eu/project/community/>) and the Galaxy community (<https://galaxyproject.org/community/>).

References

1. T. Reiter, P.T. Brooks, L. Irber, S.E.K. Joslin, C.M. Reid, C. Scott, C.T. Brown, N.T. Pierce-Ward, "Streamlining data-intensive biology with workflow systems", *GigaScience*, vol.10, no.1, pp:1-19, January 2021, <https://doi.org/10.1093/gigascience/giaa140>
2. C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, D. Garijo, Y. Gil, M.R. Crusoe, K. Peters, D. Schober, "FAIR Computational Workflows. *Data Intelligence*" vol.2, no.1, pp:108–121, 2020, https://doi.org/10.1162/dint_a_00033
3. T. Kluyver, B. Ragan-Kelley, F. Pérez, B.E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J.B. Hamrick, J. Grout, S. Corlay et al "Jupyter notebooks—a publishing format for reproducible computational workflows" In F Loizides, B Schmidt (eds) *International conference on electronic publishing*. IOS Press, ELPUB, Göttingen, 2016, pp:87–90
4. P. Di Tommaso, M. Chatzou, E. Floden, P.P. Barja, E. Palumbo, C. Notredame, "Nextflow enables reproducible computational workflows". *Nat Biotechnol* vol.35, pp:316–319, 2017, <https://doi.org/10.1038/nbt.3820>
5. J. Köster, S. Rahmann, "Snakemake—a scalable bioinformatics workflow engine", *Bioinformatics*, vol.28, no.19, pp:2520–2522, October 2012, <https://doi.org/10.1093/bioinformatics/bts480>
6. E Afgan, D. Baker, B Batut, et al. (2018) "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update", *Nucleic Acids Research*, vol.46, pp:W537–W544, 2018, <https://doi.org/10.1093/nar/gky379>
7. M.R. Crusoe, S. Abeln, A. Iosup, P. Amstutz, J. Chilton, N. Tijanić, H. Ménager, S. Soiland-Reyes, B. Gavrilović, C. Goble, "The CWL Community Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language", *CACM*, vol.65, no.6, pp:54-63 June 2022, <https://doi.org/10.1145/3486897>
8. F. da Veiga Leprevost et al, BioContainers: an open-source and community-driven framework for software standardization, *Bioinformatics*, vol.33, no.16, pp: 2580–2582, August 2017, <https://doi.org/10.1093/bioinformatics/btx192>
9. J. Ison, et al. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Research*. 2015, vol.44, no.D1, pp:D38–D47 January 2016, <https://doi.org/10.1093/nar/gkv1116>
10. C. Goble, S. Soiland-Reyes, F. Bacall, S. Owen, L. Pireddu, S. Leo. EOSC-Life Implementation of a mechanism for publishing and sharing workflows across instances of the environment. 2023, Zenodo. <https://doi.org/10.5281/zenodo.7886545>
11. M. Barker, N.P. Chue Hong, D.S. Katz, A-L. Lamprecht, C. Martinez-Ortiz, F. Psomopoulos, J. Harrow, L.J. Castro, M. Gruenpeter, P. Andrea Martinez, T. Honeyman. "Introducing the FAIR Principles for research software". *Sci Data* 9, vol.622, 2022, <https://doi.org/10.1038/s41597-022-01710-x>
12. A. Gray, L.J. Castro, N. Juty, C. Goble "Schema.org for Scientific Data" in A. Choudhary, G. Fox, T. Hey (eds) *Artificial Intelligence for Science*, pp:495-514, 2023, https://doi.org/10.1142/9789811265679_0027
13. J. Ison, M. Kalas, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S. Pettifer, P. Rice, "EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats", *Bioinformatics*, vol.29, no.10, pp:1325-32, May 2013 <https://doi.org/10.1093/bioinformatics/btt113>
14. S. Soiland-Reyes, P. Sefton, M. Crosas, L.J. Castro, F. Coppens, J.M. Fernández, D. Garijo, B. Grüning, M. La Rosa, S. Leo, E. Ó Carragáin, M. Portier, A. Trisovic, RO-Crate Community, P. Groth, C. Goble "Packaging Research Artefacts with RO-Crate", *Data Science*, vol.5, no.2, pp: 97 – 138. 2022, <https://doi.org/10.3233/DS-210053>

15. S. Soiland-Reyes, P. Sefton, L.J. Castro, F. Coppens, D. Garijo, S. Leo, M. Portier, P. Groth, "Creating lightweight FAIR Digital Objects with RO-Crate", *Research Ideas and Outcomes* vol.8, no.e93937, 2022, <https://doi.org/10.3897/rio.8.e93937>

Harmonized research information for classifying and linking research data

Fadwa Alshawaf¹, Rolf Guescini¹, Florian Kotschka¹, Maik Bierwirth¹, and Malte Dreyer¹

¹ Humboldt University of Berlin, Germany

Motivation

A well-described, harmonized in standard formats and linked metadata accelerates and refines the search process. Hence, researchers can ensure that their data is properly organized, easily discoverable, and accessible to others. Machine-readable metadata is especially essential for automatic discovery of research datasets and outputs. It facilitates the search and retrieval of data with increased accuracy, which saves the user time in finding crucial information within their field of research. This requires rich and categorized metadata that allows computers to automatically retrieve and sort relevant search results. Well documented metadata which is harmonized, standardized and serialized in linked data formats increases not only the accuracy of the search results within a field of research, but can also provide relevant and helpful suggestions within disciplinary and interdisciplinary fields of research.

Research information platform with VIVO

Within the framework of the Berlin University Alliance, we are creating a platform for structured capturing and transparent presentation of research information using the open-source software VIVO. The platform comprises a database and a frontend. The database of research information is linked to expert portfolios to showcase research expertise and activities, improve the discoverability of expertise, connect researchers to their work across disciplines and boundaries, as well as facilitate new research collaborations.

Since the project serves different institutions and collects data across disciplines, it is of significantly important to standardize the format of the collected research information. To achieve this, we represent and serialize the research information according to standardized linked data formats. We also classify the data under vocabularies such as those suggested with the project of interdisciplinary Research Core Dataset (Kerndatensatzforschung, KDSF) and standard vocabularies such EUDAT-B2Find and DESTATIS-subjects. This creates a comprehensive information model for heterogeneous scientific systems and supports the inter-change of research information within the Alliance.

Metadata extraction and classification using NLP

As part of the project, a faster, more cost-effective approach is developed to automatically extract and analyze research information from websites or files of research outputs. Machine learning Natural Language Processing (NLP) technologies make it possible to analyze huge amounts of texts on web pages or in documents and automatically extract research information. Texts are automatically scanned and analyzed to extract entities such as names and

organizations and specific information such as research subject and area or predefined keywords and vocabularies.

Text classification enables automatic categorization of the extracted information under predefined tags or groupings. It also allows texts to be categorized by their context without predefined categories being explicitly present in the text. Still, the use of keywords and controlled vocabularies as part of the metadata is essential to classify the research data under subject-relevant categories, improve, and fasten the search results. Therefore, it is recommended to set of predefined categories and vocabularies for the data classification. This includes research disciplines, subject areas, and interdisciplinary research fields. These categories are often arranged in information architectures called taxonomies, which can be further converted into machine-understandable ontologies.

The aim of creating an ontology is the development of knowledge graphs, which build the foundation for intelligent systems that can understand, interpret, categorize, and link research artefacts based on natural language. In order to achieve an automatic classification of the research data, neural networks are trained and created through machine learning approaches and applied to the text shared on websites or in documents to classify and generate the metadata. Moreover, a network of ontologies is developed to link different scientific entities such as organizations, publications, datasets, and persons to improve the discoverability of research datasets and outputs.

This approach facilitates the finding and browsing of relevant research information and quick access to research that is done at an institute under a specific discipline or area of research. It could, for example, provide necessary information for strategic planning of future development and research collaboration. It also provides relevant suggested results in the context of the searched keywords.

Outlook

We will create a platform that emphasizes research projects and links them to other research entities such as organizations, publications, datasets, funding, and events. This directs the attention to the endeavors of promoting interdisciplinary, across-institutions, and international collaborations.

Future planning within the approach of metadata classification include extended classification methods as well as ontologies. This can support broader and further classification schemes as well as multilingual search possibilities to enable a greater spectrum of findability, accessibility, interoperability and reusability of the research data.

The case for a common, reusable Knowledge Graph Infrastructure for NFDI

Lozana Rossenova¹[\[https://orcid.org/0000-0002-5190-1867\]](https://orcid.org/0000-0002-5190-1867), Moritz Schubotz²[\[https://orcid.org/0000-0001-7141-4997\]](https://orcid.org/0000-0001-7141-4997),
and Renat Shigapov³[\[https://orcid.org/0000-0002-0331-2558\]](https://orcid.org/0000-0002-0331-2558)

¹ TIB – Leibniz Information Centre for Science and Technology, Germany

² FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany

³ Universitätsbibliothek Mannheim, Germany

Keywords: NFDI, Base4NFDI, FAIR principles, research data management, knowledge graphs, knowledge graph infrastructure

Introduction

The Strategic Research and Innovation Agenda (SRIA) of the European Commission identifies *Knowledge Graphs (KGs)* as one of the most important technologies for building an interoperability framework and enabling data exchange among users across countries, sectors, and disciplines [1]. KG is a graph-structured knowledge base containing a terminology (vocabulary or ontology) and data entities interrelated via the terminology [2]. KGs are based on semantic web technologies (RDF, SPARQL, etc.) and often used for agile data integration. KGs also play an essential role within Germany as a vehicle to connect research data and research-related entities and make those accessible – examples include the GESIS Knowledge Graph Infrastructure, TIB Open Research Knowledge Graph, and GND.network. Furthermore, the Wikidata knowledge graph, maintained by Wikimedia Germany, contains a large number of research-related entities and is widely used in scientific knowledge management in addition to being an important advocacy tool for open data [3]. Extending domain-specific ontology-supported KGs with the multidisciplinary, crowdsourced knowledge in Wikidata KG would enable significant applications. The linking between expert knowledge systems and world knowledge empowers lay persons to benefit from high-quality research data and ultimately contributes to increasing confidence in scientific research in society.

Motivation

To date, several NFDI consortia have started building individual KG solutions or providing KG-compatible data using different formats and endpoints [4, 5, 6]. In addition, many of the working groups within the cross-disciplinary Sections of NFDI also include KG technologies as part of their goals and planned activities, e.g. with regards to the provision of persistent identifiers [7], ontology mappings [8], data integration and exchange endpoints [9], or training materials for RDM [10]. In some cases, individual solutions are required to meet domain-specific requirements [11]. However, in many cases, the technical and organizational overhead to run KG services can be hard to justify and is a burden to individual consortia or individual institutional members of consortia. To start discipline-specific KG work easily, NFDI consortia, participant institutions and researchers need reusable, scalable *Knowledge Graph Infrastructure (KGI)*. KGI in the context of NFDI would not only include a triplestore or graph database. KGI should encompass a whole ecosystem of software allowing to create a KG, including tools for data

import, validation and export, collaborative frontends, search APIs and SPARQL endpoints with result visualization widgets, Extract-Transform-Load and data linking software adapted to the technology stack.

Service proposal

Developed in the context of base services for the NFDI, the Knowledge Graphs Working Group from Section "Metadata, Terminologies, Provenance" is proposing a pilot KGI which provides KG infrastructure-as-a-service, combining the ease of use of software like Wikidata with research-backed data. This includes allowing NFDI stakeholders to create KGs without administrative overhead; developing an interoperability framework for connecting KGs with research infrastructures; and establishing a KGI-consultancy to increase adoption of the KGI-service.

The pilot KGI will be developed in an agile way, starting from one specific tool suite as a 'minimum viable product', and after an initialization phase will be expanded to meet the needs and requirements established through consultation with relevant NFDI stakeholder communities. The starting tool suite will be Wikibase, the open source software behind Wikidata KG, which is already being used by various consortia and participating institutions across a range of use cases. For example, MaRDI and BERD4NFDI are using Wikibase instances as central portals for all research data generated by their respective consortia participants [4, 5]. In the context of NFDI4Culture, TIB's Open Science Lab deploys Wikibase instances to structure data about digitized cultural objects entered and edited via a 3D-Viewer and Annotation tool [3]. NFDI4Memory have recently also partnered with an existing major Wikibase project, Fact-Grid, run by the Gotha Research Centre and hosted at the University of Erfurt, which provides a central repository for data about historical persons and events [12].

The growing adoption of Wikibase and the popularity of Wikidata itself – both as a repository to upload data to, and a rich resource on the linked open data (LOD) cloud to federate with – can serve as proof-of-concept that an approach to KGs involving a mix of human- and machine-readable interfaces can lower the barrier to participation across a wide range of disciplinary fields and foster the creation of complex, cross-disciplinary connections. Given the increasingly interdisciplinary use of resources (e.g., datasets, methods or models), ease-of-access to participating in the creation and exchange of FAIR research data in different fields is crucial and lays the foundation for multi- and cross-disciplinary discovery.

The identical approach to interfacing with data, modelling and querying data between Wikidata and Wikibase provides native federation capabilities across these services, which NFDI consortia that use Wikibase are taking advantage of [3, 4, 12]. Developing the Wikibase tool suite further in the context of a pilot KGI for NFDI will streamline deployment and customization, and improve interoperability at scale. This work is intended to lessen the burden on individual consortia, use existing synergies, and provide ready-to-use infrastructure for KGs.

Outlook

Beyond an initialization phase, the success of KGI in NFDI will depend on growing adoption of the services and computational methods (e.g. NLP or ML models) enabled by such an infrastructure. The interoperability framework will have to extend across Wikibase, Wikidata and other KG tool suites towards a unified NFDI with a EOSC compatibility layer. Regular consulting with and feedback-seeking from consortia partners will also be important for the expansion of the infrastructure service with relevant new offerings and for adapting to new use cases.

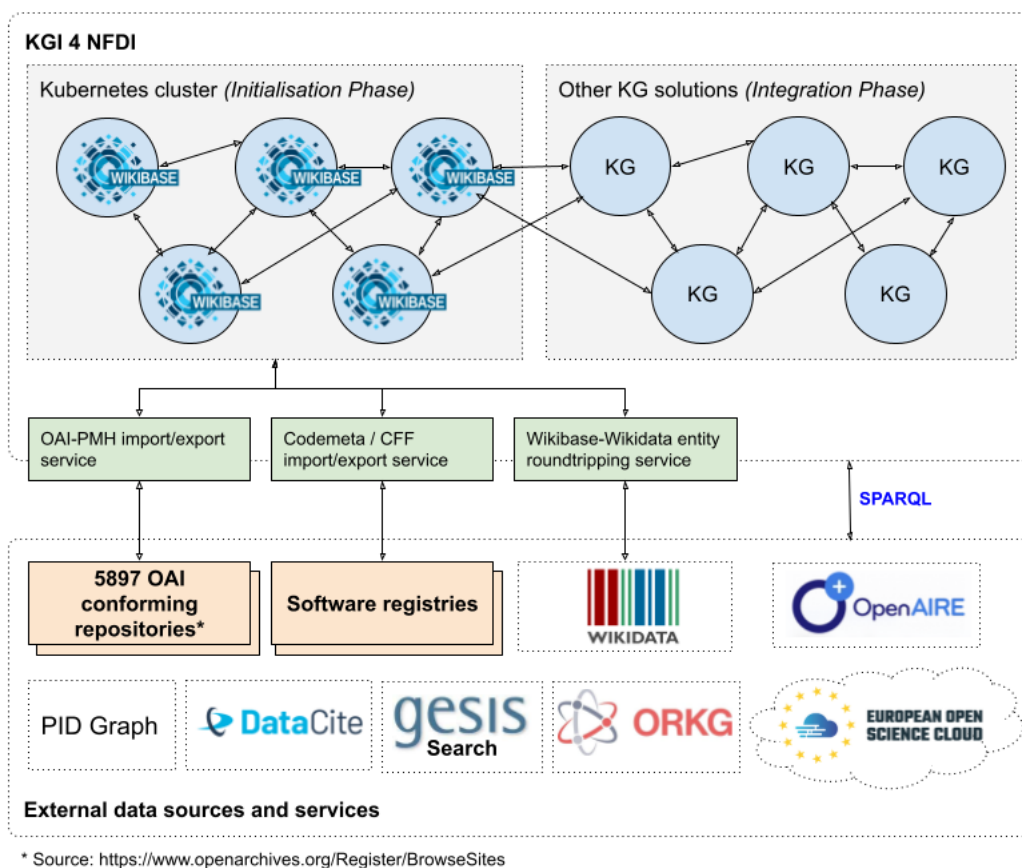


Figure 1. Overview diagram of the working concept for the KGI service.

Competing interests

The authors declare that they have no competing interests.

References

1. European Commission, Directorate-General for Research and Innovation, "Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC)," Publications Office of the European Union, 2022. [Online]. Available: <https://data.europa.eu/doi/10.2777/935288>
2. [A. Hogan, C. Gutierrez, M. Cochcz, G. de Melo, S. Kirranc, A. Pollcrs, et al, *Knowledge Graphs*. Springer Cham, 2022. [Online]. Available: <https://doi.org/10.1007/978-3-031-01918-0>
3. L. Rossenova, P. Duchesne, and I. Blümel, "Wikidata and Wikibase as complementary research data management services for cultural heritage data," in *Proc. of the 3rd Wikidata Workshop 2022, co-located with the 21st International Semantic Web Conference (ISWC2022)*, Virtual Event, Hangzhou, China, October 2022. [Online]. Available: <https://ceur-ws.org/Vol-3262/paper15.pdf>
4. The MaRDI consortium, "MaRDI: Mathematical Research Data Initiative Proposal," Zenodo, 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6552436>
5. R. Shigapov and I. Schumm. (2021). BERD: The knowledge graph of German companies. Presented at Wikibase in Knowledge Graph based Research Data Management (NFDI) Projects. [Online]. Available: <https://madoc.bib.uni-mannheim.de/58793>
6. S. Bruns, H. Fliegel, E. Posthumus, H. Sack, T. Schrade, and T. Tietz. (2023). Knowledge Graph-basierte Forschungsdatenintegration in NFDI4Culture. Presented at

- DHd2023 - Open Humanities Open Culture, Belval and Trier. [Online]. Available: <https://doi.org/10.5281/zenodo.7748740>
7. S. Bingert, J. Brase, F. Burger, B. Dreyer, S. Hagemann-Wilholt, P. Vierkant, and P. Wieder, "Concept for Setting up the Persistent Identifier Services Working Group in the NFDI Section "Common Infrastructures" (1.0)," *Zenodo*, 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6507760>
 8. I. Anders, T. Arera-Rütenik, S. Arndt, R. Baum, N. Betancort, I. Blümel, C. Busse, A. Daniel, F. Engel, L. Ghiringelli, S. Hachinger, H.r Israel, N. Karam, A. Kranz, R. Lenz, D. Linke, T. Petrenko, L. Rossenova, D. Schulz, and N. Kockmann, "Ontology Harmonization and Mapping - Working Group Charter (NFDI section-metadata) (1.0)," *Zenodo*, 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6726519>
 9. M. Stocker, L. Rossenova, R. Shigapov, N. Betancort, S. Dietze, B. Murphy, C. Bölling, M. Schubotz, and O. Koepler, "Knowledge Graphs – Working Group Charter (NFDI Section-Metadata) (1.2)." *Zenodo*, 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.7802304>
 10. P. Pelz, S. Herres-Pawlis, J. Liermann, F. Strauß, D. Ohse, N. Kockmann, M. Liebau, D. Tschink, J. Dierkes, J. Vandendorpe, R. Müller, K. Förstner, T. Hamann, C. Keßler, J. O. Heuer, D. Hausen, K. Sauerland, A. Bonn, A. Münzmay, and T. Hörner, "Working Group Charter Training Infrastructures (1.0)," *Zenodo*, 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6478698>
 11. P. Strömert and O. Koepler. (2022). Ontologies4Chem: A use case to build a NMR research data knowledge graph. Presented at ACS Spring Meeting. 23.03.2022. [Online]. Available: <https://docs.google.com/presentation/d/1qr2OiFVW4u-KFjtD71D08zOeEdH1q-nh/edit#slide=id.p1>
 12. O. Simons. "FactGrid Goes NFDI." FactGrid Blog. <https://blog.factgrid.de/archives/3104> (accessed April 20, 2023).

Organizing Scholarly Knowledge in the Open Research Knowledge Graph

An Open-Science Platform for FAIR Scholarly Knowledge

Sören Auer^{1,2}[\[https://orcid.org/0000-0002-0698-2864\]](https://orcid.org/0000-0002-0698-2864), Markus Stocker^{1,2}[\[https://orcid.org/0000-0001-5492-3212\]](https://orcid.org/0000-0001-5492-3212),
Oliver Karras¹[\[https://orcid.org/0000-0001-5336-6899\]](https://orcid.org/0000-0001-5336-6899),
Allard Oelen¹[\[https://orcid.org/0000-0001-9924-9153\]](https://orcid.org/0000-0001-9924-9153),
Jennifer D'Souza¹[\[https://orcid.org/0000-0002-6616-9509\]](https://orcid.org/0000-0002-6616-9509), and
Anna-Lena Lorenz¹[\[https://orcid.org/0000-0002-1660-1463\]](https://orcid.org/0000-0002-1660-1463)

¹TIB Leibniz Information Centre for Science and Technology, Germany

²L3S Research Center, University of Hannover, Germany

Abstract: The Open Research Knowledge Graph (ORKG) is an Open Science digital infrastructure for the production, curation, publication, and reuse of machine-actionable scholarly knowledge. Built on top of the RDF data model and extensible ontologies, the ORKG provides a common vocabulary for researchers to describe their research contributions and data, improving the discoverability and reusability of scholarly knowledge and research data. The ORKG includes tools for visualizing the relationships between different entities, making it easier to understand the connections between different pieces of research and their findings. It facilitates collaboration between researchers by providing a collaborative platform for organizing and sharing scholarly knowledge and data, reducing duplication and enabling more efficient use of resources. As research becomes increasingly data-driven, tools like the ORKG will become essential for enabling efficient, transparent, and collaborative research.

Keywords: Scholarly Communication, Knowledge Graph, Ontologies, Open-Science

1 Introduction

While many domains have significantly transformed in the digital age (e.g. encyclopedias, mail-order catalogs, street maps) scholarly communication is still based on static and relatively unstructured documents. This results in severe problems, such as a publication flood, deterioration of peer review, and inadequacy of machine assistance. The Open Research Knowledge Graph (ORKG) is a novel open-science platform that allows researchers to share and access research knowledge and data in a more efficient, transparent, and collaborative way [1], [2]. In this essay, we will explore what the ORKG is, how it works, and its potential benefits for research data stewards.

2 What is the ORKG?

The ORKG is a knowledge graph that connects scientific research papers, datasets, and methods. It is an open-science platform that provides a structured way of organizing and sharing research knowledge and data. The ORKG is designed to improve the discoverability and reusability of research data, enabling researchers to build on each other's work and collaborate more effectively. The core feature of the ORKG is to describe research contributions (e.g. published in an article) in a semantic manner. Based on these structured and semantic descriptions, the ORKG provides comparisons (i.e. overviews) on research contributions addressing the same research challenge. Based on the structured descriptions of research contributions a number of other features are provided. This includes visualizations, literature lists, and review articles comprising the structured content.

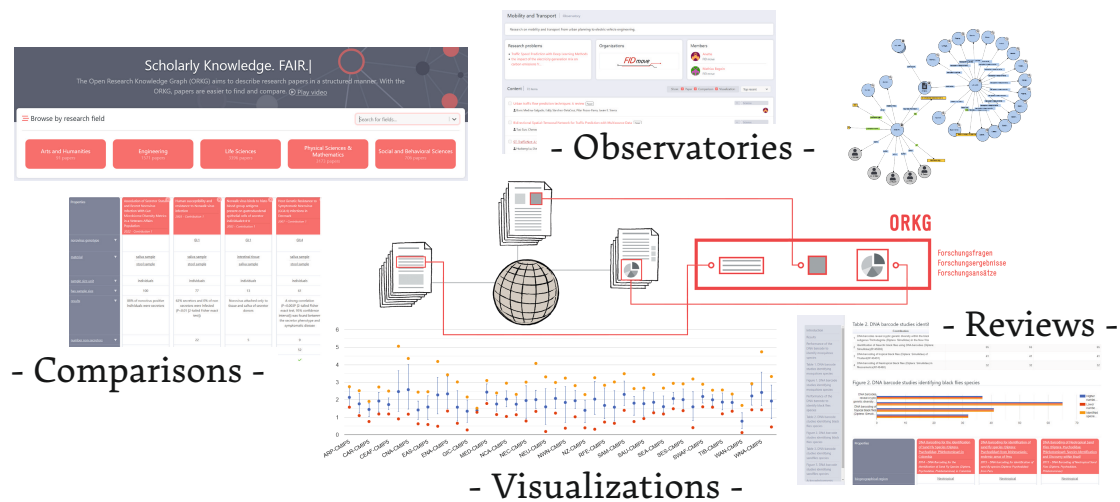


Figure 1. Semantic descriptions of research contributions can be arranged in comparisons and organized in domain-specific observatories, which in turn build the basis for visualizations, reviews, and other structured elements.

3 How does the ORKG work?

The ORKG is built on top of the Resource Description Framework (RDF) data model. Based on this data model contributors can add, curate, and organize descriptions of scientific contributions in a crowd-sourcing manner. New properties and resources for describing papers, datasets, and methods can be created on the fly, thus enabling a domain-specific and evolving ontological description of the relationships between different research entities. Pre-defined templates allow the structuring of typical, reusable knowledge patterns. The emerging domain ontologies provide a common vocabulary for researchers to describe their data, making it easier to find and understand.

The ORKG is also designed to be modular and extensible, allowing researchers to add and link to their own ontologies and customize the platform to suit their specific research needs. Automated extraction methods leveraging Natural Language Processing assist users in creating structured representations. The ORKG provides a user-friendly interface that allows researchers to create, edit, and search for research contributions.

The platform also includes tools for visualizing the relationships between different entities, making it easier to understand the connections between different pieces of research.

An example of an ORKG comparison of 25 studies regarding scenarios for the transition to renewable energy in Germany is shown in Figure 2 and available online at <https://orkg.org/comparison/R153801/>. The structured representation of the findings and scenarios from these studies enables novel ways of analysis e.g. about the installed capacity of renewable energy sources.

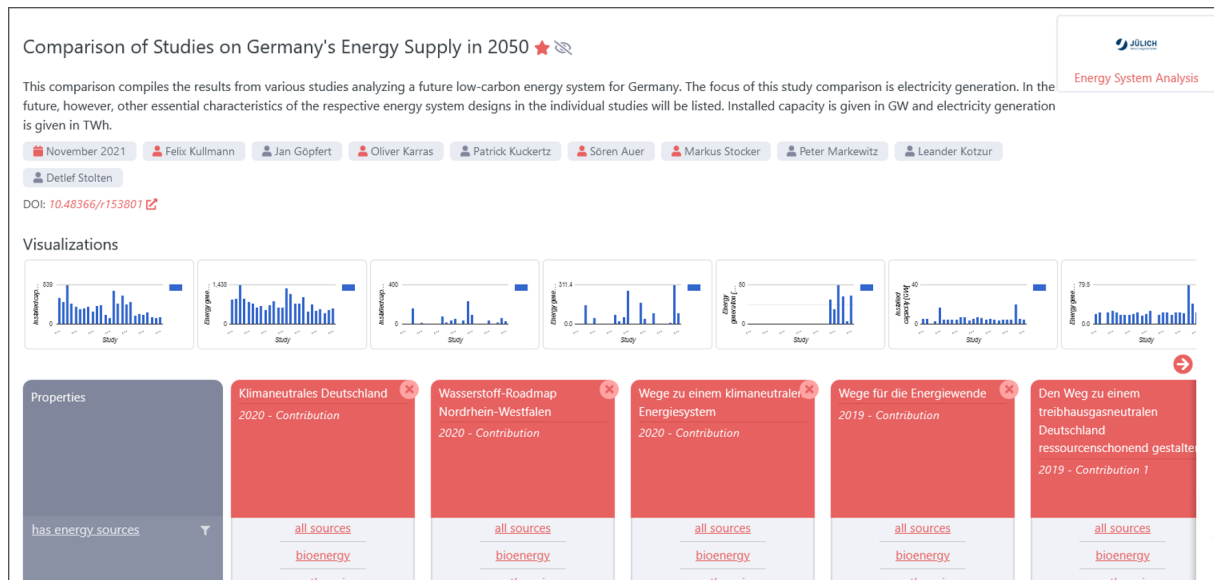


Figure 2. ORKG comparison of 25 studies regarding scenarios for the transition to renewable energy in Germany.

4 Benefits of the ORKG

The ORKG has several potential benefits for researchers. Firstly, it can improve the *discoverability* of research. By organizing research contributions into a structured format and providing a common vocabulary, the ORKG makes it easier for researchers to obtain comparative overviews of research approaches addressing a common research problem and to find relevant data for their work. This can save time and effort, as researchers no longer need to sift through large amounts of unstructured data.

Secondly, the ORKG can increase the *reproducibility* and transparency of research. By providing a clear description of the relationships between different entities and artifacts, the ORKG makes it easier for researchers to understand how data was generated and how it can be used. This can improve the reproducibility of research, as other researchers can more easily replicate experiments and verify results.

Thirdly, the ORKG can facilitate *collaboration* between researchers. By providing a shared platform for organizing and sharing research conceptualizations and data, the ORKG can enable researchers to work together more effectively (e.g. simplifying the creation of meta-analyses). This can lead to new insights and discoveries that would not have been possible otherwise.

Finally, the ORKG can help to address the issue of data silos in research. Many researchers collect and analyze data in isolation, leading to duplicate efforts and wasted resources. The ORKG provides a common platform for researchers to share their con-

tributions, data, and methods, thus reducing duplication and enabling more efficient use of resources.

5 Conclusion

The Open Research Knowledge Graph (ORKG) is an innovative novel tool for organizing and sharing research data. Built on top of the RDF data model and a set of ontologies, the ORKG provides a structured way of describing the relationships between different research entities. The ORKG has several potential benefits for research data enthusiasts, including improved discoverability, increased transparency, facilitated collaboration, and reduced data silos. As research becomes increasingly data-driven, tools like the ORKG will become essential for enabling efficient, transparent, and collaborative research.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was co-funded by the European Research Council for the project ScienceGRAPH (GA ID: 819536) as well as the DFG NFDI4Ing (no. 442146713), NFDI4DataScience (no. 460234259) and NFDI4Energy (no. 501865131) projects.

Acknowledgements

The OKRG is a team effort comprising overall more than 50 people distributed worldwide working on the development of the service platform, curating content, and researching novel features.

References

- [1] S. Auer, A. Oelen, M. Haris, *et al.*, "Improving access to scientific literature with knowledge graphs," *Bibliothek Forschung und Praxis*, vol. 44, no. 3, pp. 516–529, 2020. DOI: [10.1515/bfp-2020-2042](https://doi.org/10.1515/bfp-2020-2042).
- [2] M. Stocker, A. Oelen, M. Y. Jaradeh, *et al.*, in *FAIR Connect*, B. Magagna, Ed., vol. 1, IOS Press, 2023, pp. 19–21. DOI: [10.3233/fc-221513](https://doi.org/10.3233/fc-221513).

Knowledge Graph based RDM Solutions NFDI4Culture - NFDI-MatWerk - NFDI4DataScience

Harald Sack^{1,2}[\[https://orcid.org/0000-0001-7069-9804\]](https://orcid.org/0000-0001-7069-9804), Torsten Schrade³[\[https://orcid.org/0000-0002-0953-2818\]](https://orcid.org/0000-0002-0953-2818),
Oleksandra Bruns^{1,2}[\[https://orcid.org/0000-0002-8501-6700\]](https://orcid.org/0000-0002-8501-6700), Etienne
Posthumus¹[\[https://orcid.org/0000-0002-0006-7542\]](https://orcid.org/0000-0002-0006-7542), Tabea Tietz^{1,2}[\[https://orcid.org/0000-0002-1648-1684\]](https://orcid.org/0000-0002-1648-1684),
Ebrahim Norouzi^{1,2}[\[https://orcid.org/xxx\]](https://orcid.org/xxx), Jörg Waitelonis¹[\[https://orcid.org/0000-0001-7192-7143\]](https://orcid.org/0000-0001-7192-7143), Heike
Fliegl¹[\[https://orcid.org/0000-0002-7541-115X\]](https://orcid.org/0000-0002-7541-115X), Linnaea Söhn³[\[https://orcid.org/0000-0001-8341-1187\]](https://orcid.org/0000-0001-8341-1187), Julia
Tolksdorf³[\[https://orcid.org/0000-0002-0495-5897\]](https://orcid.org/0000-0002-0495-5897), Jonatan Jalle Steller³[\[https://orcid.org/0000-0002-5101-5275\]](https://orcid.org/0000-0002-5101-5275),
Abril Azócar Guzmán⁴[\[https://orcid.org/0000-0001-7564-7990\]](https://orcid.org/0000-0001-7564-7990), Said
Fathalla⁴[\[https://orcid.org/0000-0002-2818-5890\]](https://orcid.org/0000-0002-2818-5890), Ahmad Zainul Ihsan⁴[\[https://orcid.org/0000-0002-1008-4530\]](https://orcid.org/0000-0002-1008-4530),
Volker Hofmann⁴[\[https://orcid.org/0000-0002-5149-603X\]](https://orcid.org/0000-0002-5149-603X), Stefan
Sandfeld⁴[\[https://orcid.org/0000-0001-9560-4728\]](https://orcid.org/0000-0001-9560-4728), Felix Fritzen⁵[\[https://orcid.org/0000-0003-4926-0068\]](https://orcid.org/0000-0003-4926-0068), Amir
Laadhar⁵[\[https://orcid.org/0000-0001-9106-8825\]](https://orcid.org/0000-0001-9106-8825), Sonja Schimmler⁶[\[https://orcid.org/0000-0002-8786-7250\]](https://orcid.org/0000-0002-8786-7250), and
Peter Mutschke⁷[\[https://orcid.org/0000-0003-3517-8071\]](https://orcid.org/0000-0003-3517-8071)

¹FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

²Karlsruhe Institute of Technology (AIFB), Kaiserstr. 89, 76133 Karlsruhe, Germany

³Academy of Sciences and Literature Mainz, Geschwister-Scholl-Straße 2, 55131 Mainz, Germany

⁴Institute for Advanced Simulations – Materials Data Science and Informatics (IAS-9), Forschungszentrum Jülich GmbH, Germany

⁵University of Stuttgart, Cluster of Excellence SimTec, Stuttgart Center for Simulation Science, Universitätsstraße 32 70569 Stuttgart, Germany

⁶Fraunhofer Institute for Open Communication Systems (FOKUS), Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany

⁷GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Köln, Germany

Abstract: Based on our experience within the NFDI4Culture and NFDI-MatWerk projects we propose generalized knowledge graph based research data management solutions, which are applicable to other consortia. Our solution covers the construction of a common NFDI core ontology adapted to specific domains via domain extensions as a basis for a knowledge graph (KG) providing information about a consortium and its related research data and software resources. This KG serves as a backend for the web portal that enables interactive access and management of this data. Already implemented for NFDI4Culture and to be adapted by NFDI-MatWerk, this solution might serve as an example solution also for other consortia. We are synchronizing our efforts with ongoing work to implement knowledge graph based research data management in NFDI4DataScience.

Keywords: Semantics, Knowledge Graphs, Ontology

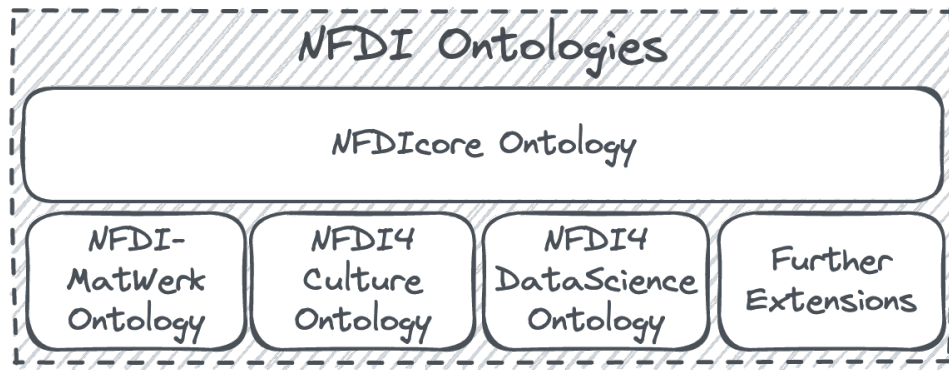


Figure 1. Schematic overview of the anticipated modular approach covering the NFDI core ontology and domain specific modular extensions.

1 NFDI Ontologies

Despite covering different scientific domains all NFDI consortia share similar concepts such as structure, organization, people, institutions, areas of expertise, data repositories, devices, infrastructure and much more that have to be represented when aiming at a semantic description [1]–[3]. This suggested the development of an NFDI core ontology to increase the interoperability of concepts across multiple NFDI consortia in the best sense of the FAIR Data Principles. A starting point was the NFDI4Culture ontology. However, the adoption by NFDI-MatWerk as a second consortium required several specific extensions, which made us realize that a modular approach works best for taking domain specific requirements into account. Figure 1 gives a schematic overview of the NFDIcore ontology, covering a consortium wide shared structure and domain specific extensions (modules) like NFDI4Culture (cto)¹, NFDI-MatWerk (mwo)² that are only relevant for a specific domain, and the NFDI4DataScience ontology that has a more interdisciplinary focus.

NFDIcore version 1.1 consists of 36 classes and 60 object attributes³. The NFDIcore classes have been linked to 24 existing ontologies, including frapo⁴, fabio⁵, void⁶ and schema⁷, following best practices in ontology development to ensure high semantic expressivity and interoperability. The ontology has been publicly available since June 2022, fully documented and integrated into the NFDI4Culture Information Portal⁸.

2 The NFDI4Culture and NFDI-MatWerk Knowledge Graphs

We distinguish between two types of NFDI Knowledge Graphs (KGs): the Research Information Graph (RIG), covering metadata about the consortium’s resources, persons, and organisations (aligned to the Common European Research Information Standard, CERIF⁹), and the Research Data Graph (RDG), covering content related index data from the consortium’s heterogeneous data resources. The goal of the RIG is to en-

¹<https://github.com/ISE-FIZKarlsruhe/nfdi4culture-ontology>

²<https://nfdi-matwerk.pages.rwth-aachen.de/ta-oms/mwo/doc/index.html>

³NFDIcore interactive view: <https://service.tib.eu/webvowl/#iri=https://nfdi4culture.de/ontology.ttl>

⁴<https://sparontologies.github.io/frapo/current/frapo.html>

⁵<https://sparontologies.github.io/fabio/current/fabio.html>

⁶<http://vocab.deri.ie/>

⁷<https://schema.org/>

⁸<https://nfdi4culture.de/ontology.html>

⁹<https://eurocris.org/services/main-features-cerif>

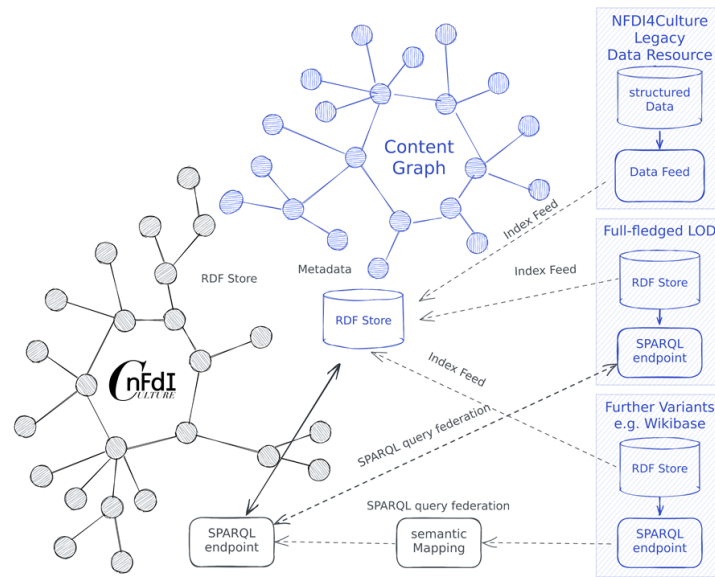


Figure 2. Overview of different data integrations. CC-BY 4.0, Authors: O. Bruns, E. Posthumus, T. Tietz, H. Sack

able exploration and retrieval of index and metadata pertaining to data resources, data services, persons, and institutions involved in NFDI consortia, while the RDG's primary objective is to facilitate access and exploration of content within the NFDI resources, as well as interconnection of the content from different resources within and across particular domains following the FAIR[4] principles. For fully fledged Linked Open Data (LOD) resources, SPARQL query federation enables access and cross connections between distributed resources. However, for reasons of efficiency and to enable across connections already locally in the RDG, the establishment of an enriched index containing metadata about entities was decided. To enable the inclusion of heterogeneous legacy data resources, including structured data and Wikibase-based resources, a metadata index harvesting mechanism has been designed based on a lightweight interchange format, the Graph Interchange Format (GIF), which so far has been implemented in a first use case as standalone protocol for NFDI4Culture as CGIF (Culture Graph Interchanged Format)¹⁰. (C)GIF can be embedded directly into any webpage to be extracted as RDF via a URL. Alternatively, data contributors can submit a (C)GIF resource in any RDF-compatible format. (C)GIF is designed to enable domain experts as contributors to easily contribute their data to the RDG without the need to implement complex APIs.

Figure 2 illustrates the concept of RIG and RDG, forming in sum the current NFDI4Culture KG¹¹ and the relevant data integration variants that are currently implemented. Further information is given in [5]¹² and [6]. The NFDI-MatWerk KG¹³ is set up in a similar manner and already allows simple retrieval via a SPARQL endpoint¹⁴. However, the (C)GIF will need to be adapted to the specific requirements of the Material Science Engineering domain.

¹⁰<https://docs.nfdi4culture.de/ta5-cgif-specification>

¹¹<https://nfdi4culture.de/resources/knowledge-graph.html>

¹²<https://doi.org/10.5281/zenodo.7748740>

¹³<https://demo.fiz-karlsruhe.de/matwerk/>

¹⁴<https://demo.fiz-karlsruhe.de/sparql/>

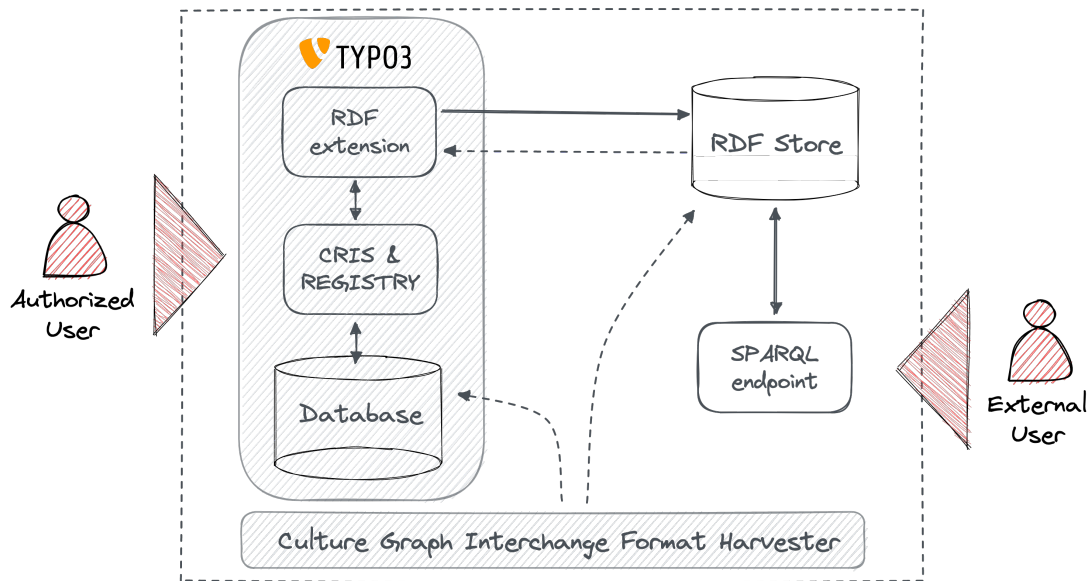


Figure 3. Overview of the Culture Information Portal Architecture. CC-BY 4.0, Author: T. Schrade

3 The NFDI4Culture Portal

NFDI4Culture is a pioneer in creating an information portal, providing a centralized access point to decentralized research data. The Culture Information Portal¹⁵ serves as a user-friendly web-based Current Research Information System (CRIS). It enables a unified access point to the research data, services, etc. of the NFDI4Culture community. The portal adheres to international standards for CRIS systems and is implemented using the TYPO3 extensions "Academy Current Research Information System"¹⁶ and "Linked Data for TYPO3"¹⁷. It allows non-expert users to contribute resources and metadata. The CRIS data is made available as Linked Open Data, with various RDF serializations and a standardized LOD API¹⁸ based on the Hydra vocabulary¹⁹. Data is curated and continuously expanded in a decentralized manner for the Culture KG using ingest routines and the Oxigraph²⁰ native RDF store. The TYPO3 implementation ensures long-term sustainability through TYPO3 LTS releases, the freedom to integrate external ontologies, and use dedicated NFDI ontologies. The establishment of the Culture Information Portal not only highlights the consortium's innovative approach but also serves as a valuable model for other consortia seeking to create similar portals and improve access to their research data.

4 Future Work

The proposed plans for future work can be divided into three main areas: Firstly, there is a need for the continued development and widespread adoption of the NFDIcore ontology and KGs for both NFDI4Culture and NFDI-MatWerk, as well as NFDI4DataScience, NFDI4Memory, and other interested consortia. Secondly, it is essential to adapt and integrate ontologies and KGs into the existing infrastructure of NFDI4DataScience.

¹⁵<https://nfdi4culture.de/>

¹⁶<https://github.com/digicademy/academy>

¹⁷<https://github.com/digicademy/lof>

¹⁸<https://nfdi4culture.de/resource/about.html>

¹⁹<https://www.hydra-cg.com/spec/latest/core/>

²⁰<https://github.com/oxigraph/oxigraph>

Lastly, a federated approach to accessing research data across multiple consortia is also being considered.

As the development of the proposed KG-based RDM infrastructure is still in its early phase, the current focus is on further developing the KG-based infrastructure and (C)GIF exchange format in NFDI4Culture and NFDI-MatWerk. Additionally, broad adoption of (C)GIF by NFDI4Culture participants, including further data repository variants such as infrastructures based on WissKI²¹, will be fostered through hands-on tutorials within the consortium, workshops for the entire domain-specific community, and the publication of guidelines and best practices. The next step is to adopt the proposed ontology and KG architecture in NFDI4DataScience, NFDI4Memory, and beyond.

Our long term goal is to enable and explore inter domain connections of KGs across multiple consortia to enable FAIR access to research data over multiple scientific domains as one of the original visions of NFDI. The proposed NFDI core ontology is seen as a key to bring this vision a step forward.

Competing interests

The authors declare that they have no competing interests.

Funding

This joint project received funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project numbers: NFDI4Culture (441958017), NFDI-MatWerk (460247524) NFDI4DataScience (460234259).

References

- [1] I. Anders, T. Arera-Rütenik, S. Arndt, *et al.*, *Ontology Harmonization and Mapping - Working Group Charter (NFDI section-metadata)*, version 1.0, Jun. 2022. DOI: [10.5281/zenodo.6726519](https://doi.org/10.5281/zenodo.6726519). [Online]. Available: <https://doi.org/10.5281/zenodo.6726519>.
- [2] I. Anders, K. Bailly, R. Baum, *et al.*, *Terminology Services - Working Group Charter (NFDI section-metadata)*, version 1.0, Jun. 2022. DOI: [10.5281/zenodo.6759325](https://doi.org/10.5281/zenodo.6759325). [Online]. Available: <https://doi.org/10.5281/zenodo.6759325>.
- [3] M. Stocker, L. Rossenova, R. Shigapov, *et al.*, *Knowledge Graphs - Working Group Charter (NFDI section-metadata)*, version 1.2, Apr. 2023. DOI: [10.5281/zenodo.7802304](https://doi.org/10.5281/zenodo.7802304). [Online]. Available: <https://doi.org/10.5281/zenodo.7802304>.
- [4] E. Schultes and P. Wittenburg, "FAIR Principles and Digital Objects: Accelerating convergence on a data infrastructure," in *Data Analytics and Management in Data Intensive Domains: 20th Int. Conference, DAMDID/RCDL 2018*, Springer, 2019, pp. 3–16.
- [5] T. Tietz, O. Bruns, H. Fliegl, E. Posthumus, T. Schrade, and H. Sack, "Knowledge Graph-basierte Forschungsdatenintegration in NFDI4Culture.," in *DHd2023: Open Humanities, Open Culture*, A. Busch and P. Trilcke, Eds., 2023.
- [6] T. Tietz, O. Bruns, L. Söhn, *et al.*, "From Floppy Disks to 5-Star LOD: FAIR Research Infrastructure for NFDI4Culture," in *3rd Workshop on Metadata and Research (objects) Management for Linked Open Science (DaMaLOS), co-located with ESWC 2023*, L. J. Castro, D. Dessi, J. Dierkes, D. Rebholz-Schuhmann, and S. Schimmler, Eds., Publisso, 2023. DOI: [10.4126/FRL01-006444986](https://doi.org/10.4126/FRL01-006444986). [Online]. Available: <https://frl.publisso.de/resource/frl:6444986>.

²¹ <https://wiss-ki.eu/>

Transparency and Involvement of Society and Policy in a Data Sharing Platform

Christina Speck¹, Patrick Jaquart¹, Christof Weinhardt¹, Johan Lilliestam², Mirko Schäfer³, Anke Weidlich³, Julia Zilles⁴, Nina Kerker⁴

¹Karlsruhe Institute of Technology, Germany

²Research Institute for Sustainability (RIFS) - Helmholtz Centre Potsdam, Germany

³Albert-Ludwigs-Universität Freiburg, Germany

⁴Soziologisches Forschungsinstitut Göttingen, Germany

Abstract:

Keywords: Data sharing platform, Energy transition, Energy policy, Public acceptance, Digital citizen science

1 Introduction and Motivation

Today, energy system models are becoming increasingly powerful and detailed regarding techno-economic parameters [1]. However, current models rarely include social and political factors, although these factors constitute important determinants for the design of energy systems [2]. For instance, any power system modelling remains irrelevant if public opposition prevents the construction of wind farms or power lines. Availability and accessibility of robust data on social and political factors are essential for policy-relevant energy modelling, but, as of now, data regarding these factors is scarce. In addition, the decentralized character of the energy transformation makes the local level increasingly important. Hence, integrating qualitative and quantitative data of the decentralised energy transformation (e.g., aspects of acceptance) is imperative for policy-relevant system modelling.

In Task Area 2 (TA2) of the nfdi4energy research project, we explore social and political drivers and constraints of the energy transition, generate and link the relevant data, and prepare it for incorporation on a data sharing platform. The aim of this task area is to co-design a scientific energy data and research sharing platform that can feed into new or existing energy models to help inform the public and political decision makers to determine socially acceptable energy pathways of the future. In addition, we will involve citizens during the project lifetime in the development of a platform that enables and incentivizes the active participation of public stakeholders in energy system research. Consequentially, the intention of this abstract within the “Linking RDM Track” is to provide an overview of the platform engagement design process for society and policy. Furthermore, we aim to discuss potential challenges with the academic audience at CoRDI 2023.

2 Task Area Objectives and Procedures

Our work in TA2 of the nfdi4energy project consists of six measures that we classify into two dimensions, depicted in Figure 1. One dimension represents the involvement of society or policy. The other dimension focuses on the measure goal of either enhancing energy system modelling with societal and political factors or engaging the public in interacting with the energy data sharing platform.

First (1), we identify future users of the nfdi4energy platform and compile a list and graphical overview of all relevant stakeholders. Additionally, we aim to identify local social drivers and constraints for the energy transition. To this end, insights into societal attitudes, perceptions, and desires regarding energy as a public good will be generated through three qualitative case studies (using a most different case design). We will conduct interviews, focus groups and workshops to examine the perspective of citizens and civil society actors in more detail and continuously integrate their perspective into the platform design process throughout the project.

Second (2), to ensure that energy system models represent a feasible option space for the future, we collect relevant regulatory data for the energy sector (i.e., landscape-related, environment-related and economic regulations). Moreover, we identify different energy and climate policy logics of past governments and large opposition parties to link them to technological change and public acceptance for Germany and other European countries.

Third (3), we identify factors determining social acceptance, support and opposition of the energy transition as well as visions for the future. Through a structured review, we identify and compile existing findings and empirical data on public sentiments regarding energy and the energy transition (e.g., [3]). The empirical and regulatory data sets are assessed with the gathered regulatory data in order to find factors determining social acceptance. In addition, the empirical data collected will be used to provide insights into future behaviors and social trends.

Fourth (4), we facilitate holistic energy system modelling with an accessible, standardized and extensive registry of modelling concepts and parameter estimations. Hence, we identify general concepts and guidelines for representing societal and political factors into energy system models. Furthermore, we collect, standardize, and integrate corresponding data and methods into the data sharing platform. Therefore we examine the status quo of incorporating societal and political factors into energy system models and develop guidelines and data sources for future energy research.

Fifth (5), we aim to motivate the public and society to share their data on the energy data sharing platform by incorporating gamification elements. Therefore, we identify use cases for gamification features, which we later implement and evaluate through user experiments.

Sixth and last (6), we identify and operationalize the most adequate visualization tool for conveying model results to stakeholders from society and politics. With a structured review, we identify best practices and requirements for scientific result communication through interactive visualisation (e.g., [4]). The most promising of the identified approaches is prototypically implemented and evaluated in a lab or online experiment.

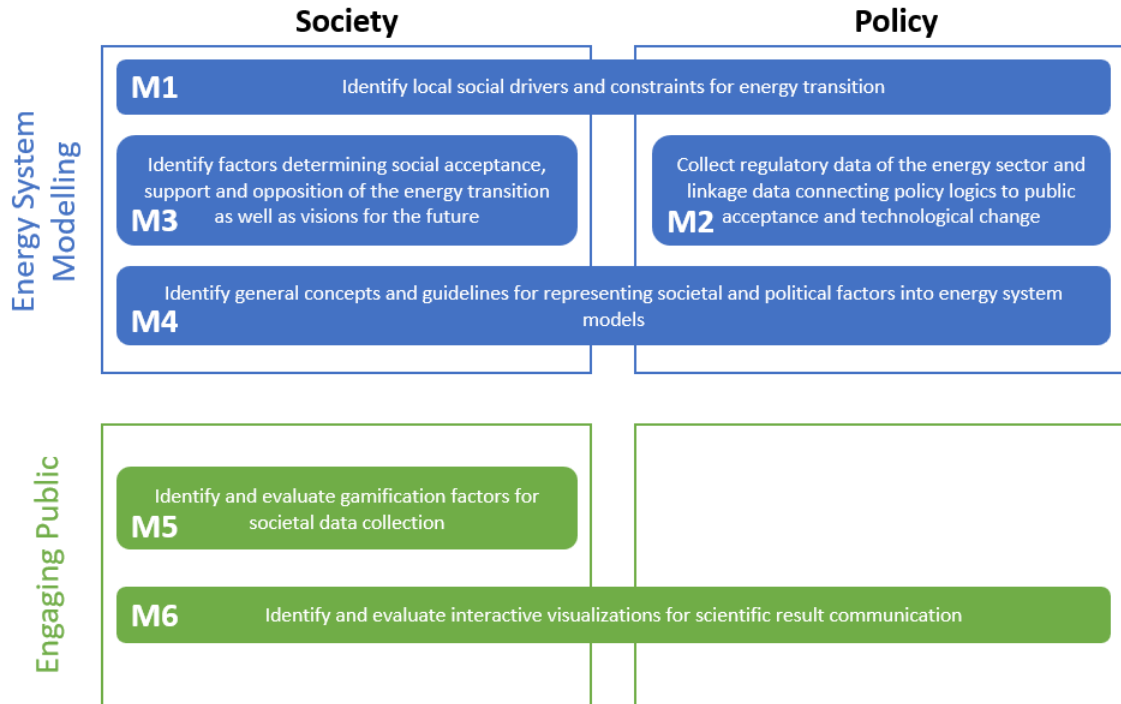


Figure 1. Classification of measures enabling transparency and involvement of society and policy in a data sharing platform.

3 Conclusion

In conclusion, our work contributes to the inclusion of public interests in the conceptualisation of energy data sharing platforms. To ensure that our results are considered in the development process of the energy data and research sharing platform, we stay in close communication with other task areas during the project time span.

Author contributions

Conceptualization, C.S., P.J.; methodology, C.S. writing—original draft preparation, C.S.; writing—review and editing, C.S., P.J., J.L., M.S., J.Z., N.K.; supervision, C.W., A.W.

Funding

The authors would like to thank the German Federal Government, the German State Governments, and the Joint Science Conference (GWK) for their funding and support as part of the NFDI4Energy consortium. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 501865131.

Competing interests

The authors declare that they have no competing interests.

ORCID iDs

Christina Speck <https://orcid.org/0009-0004-7386-8334> Patrick Jaquart <https://orcid.org/0000-0001-6995-9437> Johan Lilliestam <https://orcid.org/0000-0001-6913-5956> Mirko Schäfer <https://orcid.org/0000-0002-8029-949X> Anke Weidlich <https://orcid.org/0000-0003-2361-0912> Christof Weinhardt <https://orcid.org/0000-0002-7945-4077> Julia Zilles <https://orcid.org/0000-0002-5472-7606> Nina Kerker <https://orcid.org/0000-0003-1584-3089>

References

- [1] H.-K. Ringkjøb, P. M. Haugan, and I. M. Solbrekke, "A review of modelling tools for energy and electricity systems with large shares of variable renewables," *Renewable and Sustainable Energy Reviews*, vol. 96, pp. 440–459, 2018, ISSN: 1364-0321. DOI: [10.1016/j.rser.2018.08.002](https://doi.org/10.1016/j.rser.2018.08.002). [Online]. Available: https://www.researchgate.net/publication/327228646_A_review_of_modelling_tools_for_energy_and_electricity_systems_with_large_shares_of_variable_renewables.
- [2] A. Krumm, D. Süsser, and P. Blechinger, "Modelling social aspects of the energy transition: What is the current representation of social factors in energy models?" *Energy*, vol. 239, p. 121 706, 2022. DOI: [10.1016/j.energy.2021.121706](https://doi.org/10.1016/j.energy.2021.121706).
- [3] Research Institute for Sustainability, *Social sustainability barometer of the energy transition*, 19.04.2023. [Online]. Available: <https://www.rifs-potsdam.de/en/barometer>.
- [4] Institute for Information Management in Engineering, *View-bw - visualisierung der energiewende in baden-württemberg*, 19.04.2023. [Online]. Available: https://www.imi.kit.edu/english/46_3403.php.

Data and Services for Spatial Sustainability Science

The Story of the new IOER Research Data Centre

Ramona Voshage¹[\[https://orcid.org/0009-0005-4670-3750\]](https://orcid.org/0009-0005-4670-3750), Sujit Kumar Sikder¹[\[https://orcid.org/0000-0002-0265-7394\]](https://orcid.org/0000-0002-0265-7394),
Stefano Della Chiesa¹[\[https://orcid.org/0000-0002-6693-2199\]](https://orcid.org/0000-0002-6693-2199), Tobias Krüger¹[\[https://orcid.org/0000-0002-7085-8155\]](https://orcid.org/0000-0002-7085-8155),
Martin Schorcht¹[\[https://orcid.org/0000-0002-9898-2975\]](https://orcid.org/0000-0002-9898-2975), Gotthard Meinel¹[\[https://orcid.org/0000-0002-9201-7664\]](https://orcid.org/0000-0002-9201-7664)

¹ Leibniz Institute of Ecological Urban and Regional Development, Dresden, DE

Abstract. The emerging research data centre (RDC) at the Leibniz Institute of Ecological Urban and Regional Development (IOER) constitutes an essential milestone towards promoting sustainable land transition and transformative urban and regional development. The IOER RDC leverages spatial data science and artificial intelligence to process and analyse diverse and complex data sources. It provides high-resolution indicator maps of land use, ecosystems, and settlement structures, as well as cross-scale and cross-disciplinary spatial analyses, modelling, and simulations. Moreover, the IOER RDC offers digital tools to support decision-making, policy planning, and sustainable transformations. Hence, the IOER RDC has the potential to foster a sustainable future by facilitating the transition towards sustainable land use and development in urban and regional areas. The IOER RDC's endeavours offer a path towards addressing pressing societal challenges, such as rapid urbanisation, environmental degradation, climate change, and social inequality.

Keywords: Research Data Centre, Spatial Data Science, Sustainable Transformation

Transformative RDC for urban and regional development

Establishing the Leibniz Institute for Ecological Urban and Regional Development (IOER) new research data centre (RDC) is a significant milestone towards supporting sustainable land transition and transformative urban and regional development. The starting point of the IOER RDC is the RatSWD-accredited geospatial research data infrastructure "Monitor of Settlement and Open Space Development (IOER Monitor)", which has been operating since 2009 [1]. The IOER RDC aims to provide essential spatial data, analysis, and digital tools that enable interdisciplinary research, support policy and planning practices, and aid decision-making for spatial sustainability transformations to happen [2], [3], [4]. The IOER RDC activities respond to pressing societal challenges, including rapid urbanisation [2], environmental degradation [3], climate change [4], and social inequality [5]. To address these issues, the IOER RDC focuses on spatial data science and artificial intelligence to process and analyse heterogeneous data sources, make sense of complex spatial relationships and dynamics, and visualise the results in an accessible way. The centre provides high-resolution indicator maps on land use, ecosystems, settlement structures, building stocks, cross-scale and cross-disciplinary spatial analyses, modelling, and simulations. The IOER's RDC is committed to achieving its goals by reconstructing historical developments, describing the current status quo, and presenting alternative spatial scenarios to evaluate possible future development paths. Additionally, the IOER RDC will also develop and provide digital tools that support decision-making for sustainable transformations. These tools should help policymakers and planners understand the po-

tential impacts of different development scenarios, identify trade-offs, and make informed decisions that support sustainable urban and regional development. Data literacy is paramount in contemporary society, and the IOER RDC recognises its significance. The centre aims to improve data literacy by making data and information more accessible and comprehensible. The centre will organise regular training and collaborative educational events, workshops, and webinars to enhance the thematic data literacy of researchers, policymakers, and planning practitioners. Moreover, interdisciplinary cooperation and knowledge exchange will be facilitated between research areas, political decision-makers, and planning practitioners. The IOER RDC actively participates in national and international research data infrastructure initiatives, including NFDI4Biodiversity, NFDI4Earth, KonsortSWD, BERD@NFDI and NFDI4memory. The centre's involvement in these initiatives includes contributing to pilot projects, case studies and incubators to develop innovative solutions and data products. Its involvement is expected to generate enduring impacts in developing a robust and sustainable research data infrastructure ecosystem that will benefit various stakeholders. The IOER RDC's interdisciplinary approach focuses on spatial data science and dedication to improving competence in understanding and performing knowledge generation during the societal shift to digital culture. Overall, establishing the IOER RDC represents a significant advance towards helping sustainable land use transition and transformation in urban and regional areas and a positive step towards building a sustainable future.

Data availability statement

This submission is not based on data or any related material.

Author contributions

Ramona Voshage: Conceptualisation, Writing – review & editing, Project administration

Sujit Kumar Sikder: Writing – review & editing

Stefano Della Chiesa: Writing – original draft, Writing – review & editing

Tobias Krüger: Writing – review & editing

Martin Schorcht: Writing – review & editing

Gotthard Meinel: Conceptualisation, Funding acquisition, Project administration

Competing interests

The authors declare that they have no competing interests.

References

1. Meinel, G., Sikder, S. K., & Krueger, T. (2022). IOER monitor: a spatio-temporal research data infrastructure on settlement and open space development in Germany. *Jahrbücher für Nationalökonomie und Statistik*, 242(1), 159-170. <https://doi.org/10.1515/jbnst-2021-0009>
2. Behnisch, M., Krüger, T., Jaeger, J. Rapid rise in urban sprawl: Global hotspots and trends since 1990 In: *PLOS Sustainability and Transformation* 1 (2022) 11: e0000034 doi.org/10.1371/journal.pstr.0000034

3. Blechschmidt, J., Meinel, G. Vergleichende Untersuchung zur Erhebung der »Tatsächlichen Nutzung« in ALKIS und der daraus abgeleiteten Zeitreihe zur Flächenneuanspruchnahme. In: *zfv – Zeitschrift für Geodäsie, Geoinformation und Landmanagement* 147 (2022) 4/2022, S.250-260. doi.org/10.12902/zfv-0400-2022
4. Behnisch, M., Hladik D., Münzinger, M., Poglitsch, H. Auf dem Weg zur klimaneutralen Stadt 2030 – Quantifizierung des urbanen Solarpotenzials der Landeshauptstadt Dresden. In: Meinel, Gotthard; Krüger, Tobias; Behnisch, Martin; Ehrhardt, Denise (Hrsg.): *Flächennutzungsmonitoring XIV: Beiträge zu Flächenmanagement, Daten, Methoden und Analysen*. Berlin: Rhombos-Verlag, 2022, (IÖR-Schriften; 80), S.239-249. doi.org/10.26084/14dfns-p024
5. de Castro Mazarro A., Sikder S. K., Aguiar Pedro, A., Spatialising inequality across residential built-up types: A relational geography of urban density in São Paulo, Brazil., *Habitat International*, Volume 119, 2022. <https://doi.org/10.1016/j.habitatint.2021.102472>.

NFDI4Energy Case-Study: Comparative Analysis and Visualisation of Long-Term Energy System Scenarios

Mirko Schäfer¹, Ramiz Qussous¹, Ludwig Hülk², Johan Lilliestam³, Anke Weidlich¹

¹INATECH, University of Freiburg, Germany

²Reiner Lemoine Institut, Berlin, Germany

³Research Institute for Sustainability (RIFS) - Helmholtz Centre Potsdam, Germany

Abstract:

Keywords: Energy system, Scenario comparison, Energy transition, Energy policy

1 Introduction and Motivation

Analysis and comparison of energy system scenarios provide valuable insights into potential transformation pathways. These studies on long-term developments can serve as new inputs for scientific research and decision-making processes, providing policymakers and other stakeholders with the necessary guidance to achieve sustainable energy systems. Generally, such scenarios are derived from energy system models which often seek a cost-optimal system design under a variety of boundary conditions, ranging from technical constraints to limits of land availability or a cap on overall greenhouse gas emissions [1]. For Germany, several larger energy system scenario studies have been published, addressing the goal of carbon neutrality in 2045 as prescribed in the German climate protection act [2]. These studies show differences in their specific methodology, sector representation, parameter settings or, more generally, overall scenario narratives. This diversity represents a challenge regarding the comparability of these studies, and consequently the ability to identify consensus and controversies in their findings. Often only limited access to data for parameter settings and scenario results is provided. Almost always the data is presented in different detail and formats, thus imposing further barriers for comparison and usability for the scientific community [3].

As one of the three use cases applied in Task Area 6 of the NFDI4Energy research project, we aim to address this challenge by providing transparent and open comparative information and data on long-term energy system scenarios. Selected scenarios for the transition towards a climate-neutral Germany will be annotated with terms from the Open Energy Ontology (OEO) [4]. The comparison is building on an already existing database infrastructure from the Open Energy Platform (OEP) [5]. Existing concepts for qualitative and quantitative comparisons will be used and improved to cover a wide range of existing energy system studies.

2 Task Area Objectives and Procedures

The Task Area consists of four measures, which are in various ways connected to other measures from different Task Areas of the NFDI4Energy project. In the first measure, the scope and requirements of the scenario analysis is defined. For this purpose, suitable scenarios and comparable parameters have to be identified and analysed. An overview regarding already existing scenario comparison studies and databases will be created. An important part of the process is the preparation of the needed data sets and the annotation of the parameters with corresponding ontology terms, which allows effective and semi-automated scenario comparisons.

In the second measure, the existing data infrastructure will be enhanced and additional scenarios will be implemented. The already existing concepts from previous projects like the research project SzenarienDB and the ongoing research project SIROP are examined and evaluated [6]. The existing Scenario Factsheets, a standardised energy scenario description, are improved and extended.

Whereas the first two measures focus on identifying, processing and comparing scenarios and data, the third measure (Develop and validate draft visualisations) addresses the visualisation and communication of scenario results. To clarify requirements for this process, target groups from the public will be selected based on results from Task Area 2 (Integrating Society and Policy in Energy Research). The goal is to provide visualisations with interactive elements, which address the target groups' interests and needs. Feedback cycles with the target groups and the scientific community will be implemented throughout the overall project to adapt the presentations to new scenario results or to specific interests from various stakeholders.

The fourth measure (Involve the public & decision-makers) will contain focus group processes with selected target groups to test the visualisation tools and interactive elements for scenario analysis. This process will not only allow to improve and adapt the scenario comparison service from this Task Area, but also helps to identify mismatches between the needs of users and modellers regarding the types of models and analysis applied to study long term scenarios of the energy transition [7], [8]. The central outcome of this measure will be a communication guide for modellers targeting specific audiences.

3 Conclusions

This case study from Task Area 6 of the NFDI4Energy project addresses different challenges from Research Data Management and Infrastructure. Data from a wide range of energy system scenario studies is collected, structured, and concepts for scenario comparison are developed and implemented. To facilitate the communication of the resulting analysis to a wide range of stakeholders, concepts for visualisations and interactive elements will be developed and tested with selected target groups. This process not only assures that the insights from these studies are understandable and accessible to the public, but also provides a valuable feedback cycle from the users of this service back to energy system modellers in the scientific community.

Author contributions

Conceptualization, M.S., R.Q., L.W.; methodology, M.S., L.W. writing—original draft preparation, M.S.; writing—review and editing, M.S., A.W., R.Q.,L.W.,J.L.; supervision, A.W.

Funding

The authors would like to thank the German Federal Government, the German State Governments, and the Joint Science Conference (GWK) for their funding and support as part of the NFDI4Energy consortium. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 501865131.

Competing interests

The authors declare that they have no competing interests.

ORCID iDs

Mirko Schäfer <https://orcid.org/0000-0002-8029-949X> Ramiz Qussous <https://orcid.org/0000-0002-1989-314X> Ludwig Hülk <https://orcid.org/0000-0003-4655-2321> Johan Lilliestam <https://orcid.org/0000-0001-6913-5956> Anke Weidlich <https://orcid.org/0000-0003-2361-0912>

References

- [1] S. Pfenninger, A. Hawkes, and J. Keirstead, "Energy systems modeling for twenty-first century energy challenges," *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 74–86, 2014.
- [2] M. Ragwitz, A. Weidlich, D. Biermann, *et al.*, *Szenarien für ein klimaneutrales Deutschland. Technologieumbau, Verbrauchsreduktion und Kohlenstoffmanagement*, Schriftenreihe Energiesysteme der Zukunft, München, 2023.
- [3] L. Hülk, B. Müller, M. Glauer, E. Förster, and B. Schachler, "Transparency, reproducibility, and quality of energy system analyses—a process to improve scientific work," *Energy strategy reviews*, vol. 22, pp. 264–269, 2018.
- [4] M. Booshehri, L. Emele, S. Flügel, *et al.*, "Introducing the open energy ontology: Enhancing data interpretation and interfacing in energy systems analysis," *Energy and AI*, vol. 5, p. 100 074, 2021.
- [5] K. Reder, M. Stappel, C. Hofmann, *et al.*, "Identification of user requirements for an energy scenario database," *International Journal of Sustainable Energy Planning and Management*, vol. 25, pp. 95–108, 2020.
- [6] H. Förster, M. Stappel, L. Emele, A. Siemons, and C. Winger, "Climate and energy scenario and projection comparison. Draft workflow & typology," This work was supported by grant 03EI1035A-D (SIROP) from the Federal Ministry for Economic Affairs and Climate Action (BMWK.IIC5)., Dec. 2022. DOI: [10 . 5281 / zenodo . 7456286](https://doi.org/10.5281/zenodo.7456286). [Online]. Available: <https://doi.org/10.5281/zenodo.7456286>.
- [7] D. Süsser, H. Gaschnig, A. Ceglaz, V. Stavrakas, A. Flamos, and J. Lilliestam, "Better suited or just more complex? on the fit between user needs and modeller-driven improvements of energy system models," *Energy*, vol. 239, p. 121 909, 2022.

- [8] L. Göke, J. Weibezahn, and C. von Hirschhausen, "A collective blueprint, not a crystal ball: How expectations and participation shape long-term energy scenarios," *Energy Research & Social Science*, vol. 97, p. 102 957, 2023.

Designing a Mobility Data Trustee (MDT)

Findings from a Multi-Disciplinary Analysis of Requirements of an MDT

Andreas Czech¹[\[https://orcid.org/0000-0002-6895-0606\]](https://orcid.org/0000-0002-6895-0606), Vivien Geenen¹[\[https://orcid.org/0009-0009-7899-8249\]](https://orcid.org/0009-0009-7899-8249),
Constantin Breß²[\[https://orcid.org/0000-0002-2133-1541\]](https://orcid.org/0000-0002-2133-1541), Marija Turkovic Popovski²[\[https://orcid.org/0009-0007-0400-151X\]](https://orcid.org/0009-0007-0400-151X), Peter Krauß¹[\[https://orcid.org/0000-0002-5869-352X\]](https://orcid.org/0000-0002-5869-352X), Till Riedel¹[\[https://orcid.org/0000-0003-4547-1984\]](https://orcid.org/0000-0003-4547-1984), Frank
Gauterin¹[\[https://orcid.org/0000-0002-0870-7540\]](https://orcid.org/0000-0002-0870-7540)

¹ Karlsruhe Institute of Technology, Germany

² FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Germany

Abstract. A large amount of different data is currently collected in the mobility sector. However, due to technical, legal, and economic hurdles, it cannot be made usable. The "TreuMoDa" project, in which a Mobility Data Trustee (MDT) is being designed and tested as a prototype in the autonomous driving test area in Baden-Württemberg, provides a solution to this problem. Data trustees are a pioneering option to enable cross-sectoral data exchange between different actors from industry, science, and society. We provide an insight into the concept of the MDT. Specifically, the requirements of various stakeholders from the automotive industry, software developers, infrastructure, cities, and transport companies are analyzed with regard to the expected organizational, legal, and technical functions of such an MDT. The concept of a data trustee is particularly relevant to research data infrastructures, as it can enable the flow of data from research to industry and vice versa. Our work therefore also will look at the MDT as part of a larger research data management scheme.

Keywords: Data Trustee, Mobility Data, Anonymization, Personal Data, Intermediary, Data Sharing

1. Why the MDT is needed

Data intermediaries are foreseen as enablers for data sharing by recently introduced regulations [1]. However, it is not obvious how trusted third parties can facilitate such data reuse. When looking at the application area of autonomous driving, one challenge is how to deal with personal information contained in necessary data to develop and validate core functionalities of the systems: If data collection and processing under the GDPR is already a challenge, data reuse, and sharing turns out to be an even larger challenge.

1.1 Motivational Example

Consider a simple research project that requires data exchange between two consortium partners: One partner generates camera data from a test vehicle. The other Partner needs the data set to train an object detection model for a driver assistance system that applies brakes automatically when pedestrians cross the road. The system recognizes pedestrian intention. In this case, anonymization of the data is required before data exchange because the data contains the visible faces of people. However, as we can assume, the data provider is neither an expert in anonymization (black boxes would draw the dataset useless) nor in the legal requirements for data exchange (we need to balance the risks of data subjects). The partners

would already require help from an independent body, we call such an intermediary third party that mediates between both sides a Mobility Data Trust (MDT).

1.2 Research Questions

Within our work, we address the following research questions:

(R1) What are the necessary functions an MDT needs to provide?

(R2) Which regulatory, technical, and organizational requirements need to be met by MDTs?

(R3) Are MDTs necessary when purely looking at research data infrastructures?

Based on research questions R1 and R2, we have followed an interdisciplinary requirement engineering approach to design a feasible concept of a data intermediary called the Mobility Data Trustee. Based on our learnings, we will further discuss R3 as an outlook to a future harmonized approach that fits both industrial data spaces and research data infrastructures.

2. Requirements of the MDT

We analyzed both risks and benefits from the organizational, legal, and technical perspectives based on domain-specific constraints and best practices.

2.1 Organizational

In particular, the customer requirements are distinguished depending on the data provider, the data user, or both. Both groups are interested in simple access to data, need trust in the MDT, and education and advice by the MDT in terms of law, anonymization, and the technical part of data sharing (upload, download). Especially, the last-mentioned point needs to be realized through personal contact with the MDT. Data users require applicable data for their points of interest. On the other hand, data providers expect a high level of safety for their own data. Therefore, a strong degree of anonymization is a condition for data sharing. Furthermore, data providers require insistence on control of the choice of potential data users. The data trustee is to function as a not-for-profit data mediator between those sometimes conflicting interests. To achieve a critical mass of users, its service needs to be advertised to actors from different sectors, and communication between them needs to be uncomplicated and transparent.

2.3 Regulatory

The legal requirements of data trustees as intermediaries are governed by legislation on data protection, data governance, and related fields with the main goal to facilitate data sharing. In this regard, the most relevant are the GDPR and DGA, which shape the concept of a data trustee, as well as determine this role and legal responsibilities. The legal analysis deals with the legal review of mobility data and the definition and classification of the anonymization process. This follows the determination of levels, methods, and guarantees of anonymity. Eventually, as data sharing is one of the main drivers of the whole process, the research also addresses conditions for access and transfer of data under the GDPR and DGA.

2.4 Technical

The infrastructure must be able to handle large amounts of data, be interoperable with a variety of common data formats and exchange standards, secure the data at transit, at rest, and in use, and allow for strong data governance policies to ensure secure and GDPR-compliant data management. These challenges are combined to a scalable and operable data platform.

3. Preliminary Concept of the MDT

The most important process is the exchange of personal data from a data provider with a data user as shown in Figure 1. To transfer data of data providers to the data users of their choice, it is needed to previously acquire the necessary consent and make parties familiar with the conditions of the data transfer and usage. An anonymous sample is created to allow the data user to check whether the data is useful for their application. Thereby, the process adheres to the legal requirements for lawful, transparent, and secure processing of personal data, particularly regarding the requirements under GDPR and DGA on the consent of the data subjects, facilitation of data sharing, and responsibility for ensuring legal compliance.

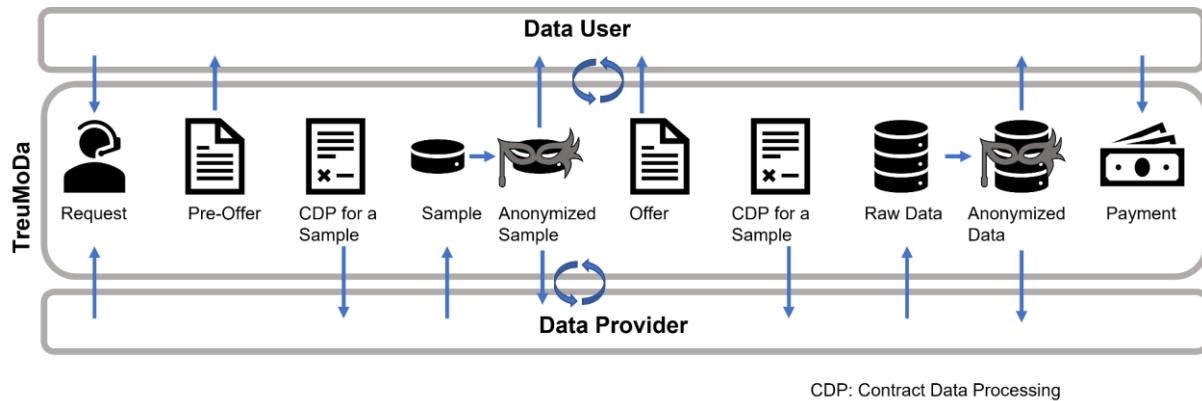


Figure 1. A simplified process of the MDT with the service of anonymization

Other developed processes are supporting data users in finding suitable data providers or checking the anonymization of the data for the current state of the art.

4. Outlook and Discussion

We are currently in the process of evaluating the conceptualized Mobility Data Trustee in the project "TreuMoDa -Treuhandstelle für Mobilitätsdaten". We believe that the concept is of interest to a national research data infrastructure as it may enable the cross-sectoral flow of data from research to industry and vice versa. Making (research) data available anonymously while retaining important aspects of the data is essential for compliance with data protection regulations and effective data reuse particularly between industry and research. The mobility domain has special requirements and faces difficult challenges regarding the handling of personal data that might not apply to all research data. Our initial analysis has shown, however, that the developed technical, legal, and organizational concepts as well as the business model and financing options are transferable. From our research, we have learned that the functions of a data trustee need to be aligned with governance and organizational structures to match regulatory requirements. Therefore, we believe it is important to consider such intermediaries early in the design of a national research infrastructure.

Data availability statement

The submission is not based on data.

Underlying and related material

There is no other material that supports our findings or is closely related to the article/contribution.

Author contributions

Andreas Czech: Methodology, Writing - Original Draft, Writing - Review & Editing **Vivien Geenen:** Visualization, Project administration, Methodology, Writing - Original Draft, Writing - Review & Editing **Constantin Breß:** Conceptualization, Investigation, Writing - Original Draft, Writing - Review & Editing **Marija Turkovic Popovski:** Investigation, Writing - Original Draft, Writing - Review & Editing **Peter Krauß:** Writing - Original Draft **Till Riedel:** Conceptualization, Methodology, Writing - Original Draft **Frank Gauterin:** Funding acquisition, Conceptualization, Supervision

Competing interests

The authors declare that they have no competing interests.

Funding

Funding for the Project "TreuMoDa - Treuhandstelle für Mobilitätsdaten" was provided by the German Federal Ministry of Education and Research (BMBF). You can find more information on the following page, among others: https://www.bildung-forschung.digital/digitalezukunft/de/technologie/daten/datentreuhandmodelle_pilotvorhaben/datentreuhandmodelle_pilotvorhaben.html

Acknowledgement

Special thanks go to Hannah Bürkle from FIZ Karlsruhe, who did various preliminary work within the framework of the responsible work package and beyond. Many thanks also to Ann-Kathrin Dreher for her work in the legal work package.

References

1. Regulation (EU) 2022/868 (Data Governance Act) [2022] OJ L 152.

Digitalizing the Chemical Landscape:

A Comprehensive Overview and Progress Report of NFDI4Chem

Oliver Koepler ¹[\[https://orcid.org/0000-0003-3385-4232\]](https://orcid.org/0000-0003-3385-4232), Christoph Steinbeck ²[\[https://orcid.org/0000-0001-6966-0814\]](https://orcid.org/0000-0001-6966-0814),
Felix Bach ³[\[https://orcid.org/0000-0002-5035-7978\]](https://orcid.org/0000-0002-5035-7978), Sonja Herres-Pawlis ⁴[\[https://orcid.org/0000-0002-4354-4353\]](https://orcid.org/0000-0002-4354-4353),
Nicole Jung ⁵[\[https://orcid.org/0000-0001-9513-2468\]](https://orcid.org/0000-0001-9513-2468), Johannes Liermann ⁶[\[https://orcid.org/0000-0003-2060-842X\]](https://orcid.org/0000-0003-2060-842X),
Steffen Neumann ⁷[\[https://orcid.org/0000-0002-7899-7192\]](https://orcid.org/0000-0002-7899-7192), Matthias Razum ³[\[https://orcid.org/0000-0002-5139-5511\]](https://orcid.org/0000-0002-5139-5511)

¹ TIB – Leibniz Information Centre for Science and Technology, Germany

² FSU - Friedrich Schiller University Jena, Germany

³ FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Germany

⁴ RWTH Aachen University, Germany

⁵ KIT - Karlsruhe Institute of Technology, Germany

⁶ JGU - Johannes Gutenberg University Mainz, Germany

⁷ IPB - Leibniz Institute of Plant Biochemistry, Halle, Germany

Abstract. The Chemistry consortium NFDI4Chem aims to digitalise key steps in chemical research, supporting scientists in managing research data throughout its life cycle. The SmartLab, embedded in a federation of services, integrates various tools such as electronic lab notebooks, data repositories, and search services, to create a smart lab environment for structured data gathering. Utilizing terminology services and adhering to data format standards, NFDI4Chem promotes secure and FAIR data sharing, fostering collaboration and expediting scientific discoveries. This development is supported by community building measures, workshops, and training initiatives, along with collaboration on international minimum information standards.

Keywords: Chemistry, Research Data, ELNs, Repositories, Ontologies, Training

1. Introduction

The Chemistry consortium NFDI4Chem envisions the digitalisation of all key steps in chemical research to support scientists in their efforts to manage research data along the data life cycle [1]. Our activities are described by the 4Chem activity clusters (Figure 1). Not all of them are directly linked to the development of a technical service, but also include measures to create a legally reliable framework of policies and guidelines for FAIR research data management (Legal4Chem), the cooperation with publishers and editors in the development of RDM author guidelines (Editors4Chem), recommendations for data and metadata standards and, the development of Minimum Information of Chemical Investigations (Standards4Chem) [2].

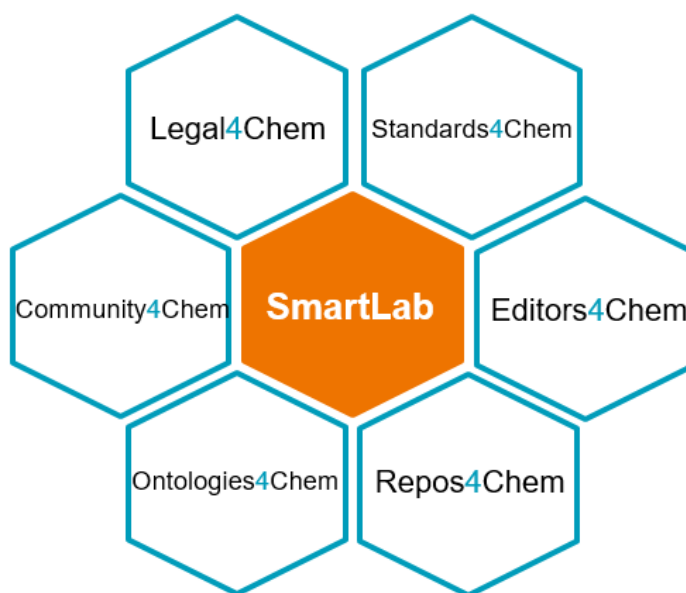


Figure 1. NFDI4Chem's Activity Clusters

2. NFDI4Chem

2.1 Terminology Service and Terminologies

The Terminology Service [3] offers a curated collection of ontologies pertinent to the chemistry community. This collection is a result of a thorough evaluation process [4]. The service enables faceted ontology searches, a granular tree and list views of classes, properties, and individuals, as well as comprehensive metadata about the ontologies. Not only does the service present an overview of ontologies in the domain, but it also strives to facilitate comparison and analysis across multiple ontologies for curation purposes. To achieve this, it provides a unified perspective on issues from the original ontology repositories within the Terminology GUI. Moreover, the terminology service offers a comprehensive API for retrieving all terminology data and information, allowing integration with other NFDI4Chem and NFDI services. These developments are accompanied by the curation and development of ontologies for the chemistry community. NFDI4Chem aims to encourage and moderate a process towards the harmonization of ontologies within chemistry. A first step was the first international Ontologies4Chem workshop with curators from all major chemistry ontologies [5]. During the analysis of the ontology landscape, gaps were identified, which are now leading to the development of new ontologies, such as the ontology for vibrational spectroscopy VIBSO. These developments are undertaken in close cooperation with the chemistry community and international standardisation bodies like the IUPAC.

2.2 Electronic Lab Notebooks (ELN) Chemotion

In a lab environment ELNs are able to collect data that is coming from devices, to process data into readable files and to manage data along with further descriptions and metadata. The more discipline specific functions are supported by an ELN, the better usually the user confidence but also the suitability of a tool with respect to the generation of FAIRdata. We provide access to different ELNs to allow a comparison of the ELNs with respect to the required functionality, and we support Chemotion ELN as a reference instance for the implementation of results and requirements gained from NFDI4Chem and the community. The Open Source ELN is constantly updated with new functions and currently more than 25 instances are installed in Germany, supporting scientists in different universities. The ELN offers different functions that are key assets for the digitalization of chemistry, in particular with respect to the work with

molecular structures. In addition, it enables a seamless data flow from a wide range of devices to the ELN and supports the transfer of data and data collections to repositories such as Chemotion repository and RADAR4Chem. In the future, Chemotion ELN should work as a flexible interface connecting the digital work environment of chemists with the federation of repositories.

2.3 Training and Education

To support and train the community on all levels, we have initiated a large bundle of services and events: first contact to NFDI4Chem can occur at the Helpdesk or at our conference booths (approx. 12 conferences per year). Together with regular newsletters as well as highly active social media accounts, this leads to wide community outreach. As a regular basis for exchange, we have also established the monthly NFDI4Chem Stammtisch with speakers from all branches of RDM, machine learning, electronic lab journals etc. Further, we offer RDM training courses in a 2-day and a condensed 1-day mode, but also Chemotion training courses and the digital Chemotion Q&A session. Best practices also help to demonstrate the application of new RDM tools to the community. Moreover, we integrated Chemotion into curricular teaching to train the next generation of chemists. As a reference module to all chemists, the NFDI4Chem knowledge base offers approved information on all topics in RDM and practical help.

2.4 Federation of Repositories

NFDI4Chem is setting up a federation of interconnected data repositories and make research data FAIR and open. This requires offering storage and long-term archival, enabling chemists to publish their data with contextual metadata and DOI. To ensure interoperability, we currently adapt metadata schemas, implement APIs, integrate Authentication and Authorisation Infrastructure (AAI) and set up a federation of interoperable services. Core repositories in NFDI4Chem are chemotion, nmrXiv, RADAR4Chem, MassBank, SupraBank, and Strenda, covering the main subdisciplines of chemistry and data types (Fig. 2).

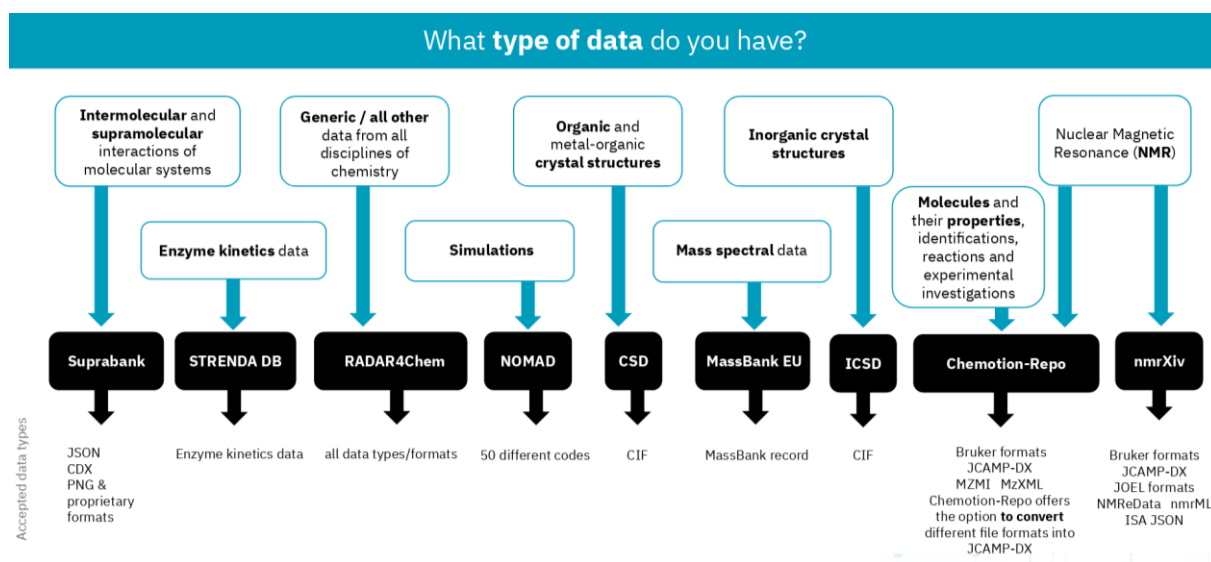


Figure 2. Data Types and Data Repositories

2.5 Search Service

The NFDI4Chem Search Service [6] harvests and indexes metadata from over 90,000 datasets provided by repositories of the NFDI4Chem federation. It provides a central access point, for instance to obtain an overview of all datasets available for specific molecular entities.

Users can filter datasets by various criteria, such as the originating repository or measurement technique. The advanced search feature allows for queries using chemical identifiers like InChI, InChI Key, and SMILES. Within the development of the search service NFDI4Chem also addresses the challenges of implementing domain specific metadata for chemical substances, chemical structural information or measurement techniques.

3. Outlook

We have now reached a point where the developed services and content have achieved a sufficient level of maturity to be rolled out to the wider community. Through our workshops, RDM trainings, and roadshows, we have built a network to reach as many chemists in the lab as possible. Initial pilot projects have been very successful. We will also increasingly focus on networking with services from consortia closely related to chemistry.

Funding

The NFDI4Chem project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the NFDI Funding Programme (grant no. 441958208).

Acknowledgement

We thank all members of the NFDI4Chem consortium, which contribute to the progress and services of NFDI4Chem.

References

1. C. Steinbeck, O. Koepler, F. Bach, S. Herres-Pawlis, N. Jung, J. Liermann, S. Neumann, M. Razum, C. Baldauf, F. Biedermann, T. Bocklitz, F. Boehm, F. Broda, P. Czodrowski, T. Engel, M. Hicks, S. Kast, C. Kettner, W. Koch, G. Lanza, A. Link, R. Mata, W. Nagel, A. Porzel, N. Schlörer, T. Schulze, H.-G. Weinig, W. Wenzel, L. Wessjohann, S. Wulle, "NFDI4Chem - Towards a National Research Data Infrastructure for Chemistry in Germany", 2020, Research Ideas and Outcomes 6, e55852, doi: [10.3897/rio.6.e55852](https://doi.org/10.3897/rio.6.e55852)
2. S. Herres-Pawlis, F. Bach, I.J. Bruno, S.J. Chalk, N. Jung, J.C. Liermann, L.R. McEwen, S. Neumann, C. Steinbeck, M. Razum, O. Koepler, "Minimum Information Standards in Chemistry: A Call for Better Research Data Management Practices", 2022, Angew. Chem. Int. Ed Engl. **61**, e202203038, doi: [10.1002/anie.202203038](https://doi.org/10.1002/anie.202203038)
3. NFDI4Chem Terminology Service, <https://terminology.nfdi4chem.de>, last accessed 28 April 2023
4. P. Strömert, J. Hunold, A. Castro, S. Neumann, O. Koepler. "Ontologies4Chem: the landscape of ontologies in chemistry", 2022, Pure Appl. Chem., doi: <https://doi.org/10.1515/pac-2021-2007>
5. P. Strömert, J. Hunold, and O. Koepler, "1st Ontologies4Chem Workshop – Ontologies for chemistry," Sep. 07, 2022, doi: [10.25798/frnp-sn04](https://doi.org/10.25798/frnp-sn04).
6. NFDI4Chem Search Service, <https://search.nfdi4chem.de>, last accessed 28 April 2023

The DAPHNE4NFDI and PUNCH4NFDI Consortia in the NFDI

L. Amelung¹[\[https://orcid.org/0000-0002-8137-1553\]](https://orcid.org/0000-0002-8137-1553), A. Barty¹,
B. Murphy²[\[https://orcid.org/0000-0002-1354-2381\]](https://orcid.org/0000-0002-1354-2381), C. Schneide¹[\[https://orcid.org/0000-0003-1024-6875\]](https://orcid.org/0000-0003-1024-6875),
A. Schneidewind³[\[https://orcid.org/0000-0002-7239-9888\]](https://orcid.org/0000-0002-7239-9888), and
T. Schoerner¹[\[https://orcid.org/0000-0002-7213-0352\]](https://orcid.org/0000-0002-7213-0352)

¹Deutsches Elektronen-Synchrotron, Hamburg, Germany

²Christian-Albrechts Universitaet zu Kiel, Germany

³Forschungszentrum Juelich, Germany

Abstract: The DAPHNE4NFDI and PUNCH4NFDI consortia represent the large scale facilities in the German physical sciences community. Work in DAPHNE4NFDI and PUNCH4NFDI is characterised by the use of large-scale research infrastructures – reactors, light sources, accelerators, telescopes, observatories, satellites – that serve international research communities of up to several thousand users and produce data in the terrabyte, often petabyte and in future exabyte range. The communities will have to master massive challenges in data management, building on and extending their leadership in “big data” management, distributed computing, multi-user management, and data loss / data irreversibility issues.

Keywords: DAPHNE4NFDI, PUNCH4NFDI

1 First section

The DAPHNE4NFDI and PUNCH4NFDI consortia represent the large scale facilities in the German physical sciences community. Together they comprise 64 co-applicant and participant institutions – universities, Helmholtz Research Centres, Leibniz Institutes, Max Planck Institutes – and close to 15,000 researchers in the related physics and natural science communities.

DAPHNE4NFDI – “DAta from PHoton and Neutron Experiments” – focuses on data management for experiments using large scale photon and neutron research facilities. The properties of photons and neutrons allow us to probe the structure of matter to find out where atoms are and how they move - in solids, liquids and thin films even at very low temperatures or high pressures, penetrate through thick materials to obtain 3D structures, and map spatial chemical distributions. This makes it possible, for example, to see the tiniest cracks and pores in a turbine blade, to find small amounts of impurities in a semiconductor, the chemical states of catalysts or batteries in operando, or to determine the overall structure of a protein molecule or virus down to position of individual atoms. Photon and neutron research therefore includes a diverse range of

scientific disciplines. Individual instruments can generate over 1PB of data per day, and each facility hosts multiple instruments. The large amount of experimental data generated at high data rates presents us with substantial challenges: The data is often user-specific, as a wide variety of software is used for experimental control, data collection and data analysis. Therefore, it is currently not easy to share the data. There is a great need for digital tools to capture the data and meta data, curate the storage and provide high-level data analysis so that the data is reusable. Significant expert knowledge is required in order to use the data. DAPHNE4NFDI addresses the research data lifecycle, aiming to make data from photon and neutron experiments accessible to non specialists thereby making scientific work more efficient and gaining more knowledge from the data collected by others.

PUNCH4NFDI – “Particles, Universe, NuClei and Hadrons” – is the NFDI consortium of particle, astro-, astroparticle, hadron and nuclear physics. PUNCH physics addresses the fundamental constituents of matter and their interactions, as well as their role for the development of the largest structures in the universe - stars and galaxies. The achievements of PUNCH science range from the discovery of the Higgs boson over the installation of a 1 cubic kilometre particle detector for neutrino detection in the antarctic ice to the detection of the quark-gluon plasma in heavy-ion collisions and the first picture ever of the black hole at the heart of the Milky Way. The prime goal of PUNCH4NFDI is the setup of a federated and FAIR science data platform, offering the infrastructures and interfaces necessary for the access to and use of data, analysis workflows and computing resources of the involved communities and beyond. PUNCH4NFDI also offers tools for the efficient scientific exploitation of research data.

Work in DAPHNE4NFDI and PUNCH4NFDI is characterised by the use of large-scale research infrastructures – reactors, light sources, accelerators, telescopes, observatories, satellites – that serve international research communities of up to several thousand users and produce data in the TB, often PB and in future EB range. New and upcoming facilities like the High-Luminosity LHC (HL-LHC), the Square Kilometre Array (SKA), the Einstein Telescope (ET), PETRA IV, or FAIR (Facility for Antiproton and Ion Research in Europe) will soon produce data with unprecedented rates, volumes and complexities, compounded by the increasing prevalence of hybrid multi-modal experiments. The communities will have to master massive challenges in data management, building on and extending their leadership in “big data” management, distributed computing, multi-user management, and data loss / data irreversibility issues. The increasing demands for FAIR and open data and science now pose additional challenges that also drive the developments in the two consortia.

There are also differences between the two consortia, stemming from the different work organisation and scientific methods involved. In particular, the separation between “facilities” and “users”: In DAPHNE4NFDI users often come to a given facility for a limited beam time, bring their individual sample, and take data at pre-installed experimental set-ups that they have not built themselves. Facility users are often expert in their own domain, but non-expert in photon science data analysis. In PUNCH4NFDI it is much more common that the scientists build their own, sometimes massive, detectors and operate it and exploit its data at the same time, sometimes over very long time-scales (several decades e.g. at the LHC), in large international collaborations that live equally long. Here, users and providers typically coincide. Nevertheless, there are also significant differences in approaches within the consortia between e.g. particle physics and astronomy, or smaller and larger enterprises.

Due to these differences, also the immediate challenges and goals of the two consortia partly differ in some aspects: For DAPHNE4NFDI, the following main tasks have been identified:

- Improve metadata capture through consistent definitions and workflows supported by user-driven online logbooks that are linked to the data collection, thus enabling a richer capture of information about the experiments than is currently possible;
- Establish a community repository of processed data, new reference databases and analysis code for published results, linked, where possible, to raw data sources, to sustainably improve access to research data and enable data and software re-use;
- Develop, curate and deploy user-developed analysis software on facility computing infrastructure so that ordinary users can benefit from and repeat the analysis performed by leading power user groups through common data analysis portals.
- Develop education and outreach programs to export the knowledge and standards developed within DAPHN4NFDI.

PUNCH4NFDI has defined the following high-level goals:

- An integrated prototype package of dynamic digital research products, a science data platform, and the Compute4PUNCH compute and storage resources (Storage4PUNCH) coupled with single-sign-on (AAI). This will, for example, enable the collaborative creation and re-use of dynamic research products including analysis and simulation workflows based on a selected range of tools and required software environments. The milestone will enable the collaboration within and among research groups on the development of analysis workflows, and the combination of research products within the supported range of formats and tools.
- Data irreversibility solutions – a new kind of science in the age of too-large-to-be-stored data streams: Irreversible data reduction and compression based on real-time decisions are a prerequisite for future discovery science in the PUNCH fields and will become more and more important in other branches of science.
- Rolling out the PUNCH4NFDI outreach and education programme: It is important to share the expertise with the wider science community.

Despite these differences, a close and lively cooperation between the two consortia exists. An assessment of synergies and the exchange of information on tools or developments takes place on regular basis.

2 Second section

Data availability statement

The submission is not based on data.

Author contributions

Conceptualization: AB, BM, AS and TS; Project Administration: LA, AB, CS and TS; Writing original draft: TS; Writing review and editing: LA, AB, BM, CS, AS and TS.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the consortia PUNCH4NFDI and DAPHNE4NFDI in the context of the work of the NFDI e.V. The consortia are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project numbers 460248186 (PUNCH4NFDI) and 460248799 (DAPHNE4NFDI).

Research Data Management for Experiments in Solid-State Physics: Concepts

Heiko B. Weber¹[\[https://orcid.org/0000-0002-6403-9022\]](https://orcid.org/0000-0002-6403-9022), Sandor Brockhauser²[\[https://orcid.org/0000-0002-9700-4803\]](https://orcid.org/0000-0002-9700-4803), Christoph Koch²[\[https://orcid.org/0000-0002-3984-1523\]](https://orcid.org/0000-0002-3984-1523),
Laurenz Rettig³[\[https://orcid.org/0000-0002-0725-6696\]](https://orcid.org/0000-0002-0725-6696), Martin Aeschlimann⁴[\[https://orcid.org/0000-0003-3413-5029\]](https://orcid.org/0000-0003-3413-5029), Walid Hetaba⁵[\[https://orcid.org/0000-0003-4728-0786\]](https://orcid.org/0000-0003-4728-0786),
Marius Grundmann⁶[\[https://orcid.org/0000-0001-7554-182X\]](https://orcid.org/0000-0001-7554-182X), Markus Kühbach²[\[https://orcid.org/0000-0002-7117-5196\]](https://orcid.org/0000-0002-7117-5196), and Michael Krieger¹[\[https://orcid.org/0000-0003-1480-9161\]](https://orcid.org/0000-0003-1480-9161)

¹Department of Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

²Department of Physics and IRIS Adlershof, Humboldt-Universität zu Berlin, Germany

³Fritz Haber Institute Berlin, Germany ⁴RPTU Kaiserslautern-Landau, Germany

⁵Max-Planck-Institute for Chemical Energy Conversion, Mülheim an der Ruhr, Germany

⁶Department of Physics, Universität Leipzig, Germany

Abstract: FAIRmat develops concepts for paving the way to enable FAIR research data in solid-state physics. For selected theoretical data in this field, the NOMAD portal has developed mature concepts and technological solutions for storing data according to the FAIR principles. Extending this approach to experimental data is challenging due to their diversity and missing standards. In this paper we present our comprehensive approach to establish FAIR data in the field of experimental solid-state physics despite its heterogeneity. The concept includes elaboration of standards, community building and methods that facilitate the community's transition to FAIR standards.

Keywords: solid-state physics, research data management, electronic lab notebook, FAIR data

1 Introduction

Within physics, solid-state physics is the largest community. It is extremely heterogeneous, spanning over small groups and large institutes with their professional infrastructure. Technically, a plethora of well-established methods, but also customized, unconventional methods are covered. Still today, metadata are often documented in handwritten laboratory notebooks, and measurement data are predominantly handled in silico, with the data-generating measurement software being often self-written. In contrast, professional stand-alone equipment is predominantly purchased with its own proprietary code, many details of which are intransparent to the user. In summary, the field is technically very advanced, but extremely heterogeneous and diverse, which imposes challenges to modern RDM. In order to enable a transition to FAIR data management in this community, FAIRmat has identified the relevant fields of action, which are presented in this manuscript.

2 Fields of Action

Laboratory Control Software. FAIRmat is currently developing a software (NOMAD CAMELS) that will make it particularly easy for scientists with setting up a new experiment and guarantees FAIR-ready data output. Communication with the instruments and the measurement protocol will be configurable in a graphical user interface. Camel's output is an HDF5 file by default, which contains not only data, but also the technical parameters of the experiment in a structured way.

Electronic Laboratory Notebooks (ELNs). ELNs play a key role in the documentation of experiments, collecting the entity of metadata. In order to enable the reusability in 'FAIR', ELNs have to become particularly simple to use at two interfaces: First, at the input stage, so that data can be entered accurately and efficiently. Second, at the services which motivate and support users with entering data sufficiently structured and schematised, so that as much automated processing becomes routinely possible. FAIRmat recognises that the community already uses different ELNs and is developing interfaces and tools to structure the data in ELNs. FAIRmat also uses the existing advantages of NOMAD Oasis and offers domain-specific data schemas via NOMAD's ELN customization capabilities, with a clear focus on structured, schematized data handling.

FAIR-ready data management. FAIRmat stresses that data shall always be collected together with metadata as these are the contextualisation which enables a proper interpretation and distilling knowledge (from the data). According to FAIR principles[1], this requires the use of vocabulary which meets community standards and so enables machine interpretability. FAIRmat recognised the requirements of our diverse community, the fast developments of cutting edge experimental techniques. For new methods, FAIRmat suggests and provides tools for a FAIR-ready documentation of all data and metadata. While these documentation may not follow widely accepted community standards, and so the data is not (yet) FAIR by definition, but is digitized and described and may hence be converted to community standards once they are developed.

Community Standards. FAIRmat encourages the different domain scientists to develop community standards required for FAIR data management. FAIRmat contributes to EMglossary harmonisation [2] work organised by Helmholtz Metadata Collaboration, and also proposes experimental-technique-specific domain ontologies for standardisation. These are developed as extensions of the NeXus community standardisation platform[3]. NeXus allows the documentation of data and metadata concepts in a structured and hierarchical way. FAIRmat is also involved in expressing the NeXus standard in the Semantic Web language OWL which supports interoperability by allowing the connection of all data and metadata to other domains or higher-level ontologies. Any experimental data collected and stored according to the NeXus standard can be automatically loaded into the NOMAD RDM solution.

Workflow in the NOMAD environment. NOMAD enables not only a safe and long-term storage of the data, but also to work with the data. FAIRmat provides examples for containerising community software solutions for data reduction, processing and refined analyzes. The containers can be launched in NOMAD to work with the data directly with no need to download or install the domain-specific tools locally but rather access your data at the server. Currently, NOMAD and its local deployment NOMAD Oasis[4] can be installed on a single server or on a Kubernetes[5] cluster to provide the required computing nodes for such remote data analysis work. All uploaded data, surplus those

data generated on the server is indexed by NOMAD and thus is available in searches to those having access rights to the given dataset. After publishing, it becomes publicly accessible and usable under the Create Commons Attribution License (cc-by) 4[6].

Instrument manufacturers as technology partners. A problem for FAIR data is that manufacturers of scientific instruments often provide software that uses proprietary formats. It is very tedious and sometimes impossible to link data and necessary metadata with the entries in the ELNs and to generate FAIR data out of proprietary sources. FAIRmat takes the approach of developing application-specific data schemes within the scientific community and presenting them to the technology partners. In joint workshops with the companies, data experts and scientists, the details are then discussed and a canon is defined for how the data should be stored in the future.

Broad data expertise. Neither FAIRmat nor local data stewards can solve the plethora of tasks that arise when our community converts all processes to FAIR data. This calls attention to the scientists themselves. In physics education, data handling is not yet considered a crucial skill. Here, a discussion was initiated within the physics community to consider data competence as an additional key competency. For example, on the initiative of FAIRmat scientists, the use of ELNs was introduced in bachelor lab courses[7]. This leads to enhanced competences of the students, and establishes ELNs as the new working standard. Moreover, lab course ELNs are a convenient sandbox in which new RDM concepts and technologies can be tried out within a time-limited framework. Only with the active cooperation of all scientists involved a sustainable transition to FAIR data can succeed.

Funding

FAIRmat is funded by the the Deutsche Forschungsgemeinschaft "DFG, German Research Foundation" – project 460197019.

References

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
- [2] *Em glossary initiative*, 2023. [Online]. Available: <https://helmholtz-metadaten.de/en/em-glossary-initiative>.
- [3] M. Könnecke, F. A. Akeroyd, H. J. Bernstein, *et al.*, "The nexus data format," *Journal of applied crystallography*, vol. 48, no. 1, pp. 301–305, 2015. DOI: [10.1107/S1600576714027575](https://doi.org/10.1107/S1600576714027575).
- [4] *Nomad - materials science data managed and shared*, 2023. [Online]. Available: <https://nomad-lab.eu/>.
- [5] *Kubernetes, also known as k8s, is an open-source system for automating deployment, scaling, and management of containerized applications*. 2023. [Online]. Available: <https://kubernetes.io/>.
- [6] *Creative commons attribution 4.0 international*, 2023. [Online]. Available: <https://creativecommons.org/licenses/by/4.0/>.
- [7] M. Krieger, H. B. Weber, and C. van Eldik, "Früh zur Datenkompetenz," *Physik Journal*, vol. 21, p. 42, 2022. [Online]. Available: <https://www.pro-physik.de/restricted-files/158142>.

Integrating Data Literacy into University Curricula

Student Centred Learning in Undergraduate Physics Lab Courses

Janice Bode¹[\[https://orcid.org/0000-0003-1777-9148\]](https://orcid.org/0000-0003-1777-9148), Philipp Jaeger²[\[https://orcid.org/0000-0002-7526-1489\]](https://orcid.org/0000-0002-7526-1489), and
Sonja Schneidewind¹[\[https://orcid.org/0000-0002-0978-1056\]](https://orcid.org/0000-0002-0978-1056)

¹University of Münster, Germany

²Pexon Consulting GmbH, Germany

Abstract: The implementation of competencies regarding Research Data Management (RDM) and Data Literacy (DL) into the curricula is a major challenge for the NFDI community. In physics as well as in other disciplines, the beginners' lab courses are the first point in time where students have to deal with data, and are therefore an intuitive point for the integration of first RDM learning objectives. However, the modification of teaching materials is a time-consuming and potentially expensive task. In this contribution the authors show how this can be achieved in the physics beginners' lab on the basis of concrete examples, thereby affecting the education of students in a wide range of subjects, including the sciences, engineering, but potentially also medicine or geology.

Keywords: Data Life Cycle, Data Literacy, Teaching Material, FAIR principles, Higher Education, Lab Courses, Graduate Studies, Physics, Sciences

1 Introduction

There are many views on how to best integrate Research Data Management (RDM) and the Research Data Life Cycle [1] as well as Data Literacy (DL) [2] in curricula across the disciplines. In this work, the authors focus on physics programmes at EQF [3] level 6 (Bachelor's degree) or combined programmes on level 6 and 7 (previous German degrees, e.g. Diplom, Staatsexamen) and programmes that contain a significant share of physics courses, such as biophysics, astronomy, material science etc.

The physics community has been discussing several approaches, including dedicated courses for RDM and DL, or practical learning during the practical phases and thesis projects of the programme. The first option suffers from the already full curriculum which causes difficulties in adding new content. Since students choose individual thesis topics, the second option does not ensure equal RDM content for all students during their studies. Furthermore, at the time of their thesis students may already have adopted some kind of handling data. A third option, which this work focusses upon, is to modify the beginners lab courses in such a way that all students learn and experience the basics of RDM and DL while taking the courses, and thereby also become familiar with the FAIR principles [4]. The key arguments in favour of such an approach are [5]:

- During the beginners lab, students work with “real” experimental data for the first time, making it a natural starting point to teach RDM.
- By sharing data during the lab, it is possible to analyse data in a more complex fashion, and hence to answer more interesting and relevant questions.
- The implementation of the necessary changes is quite straightforward. In particular, no changes to any examination regulations become necessary.
- The anticipated outcome can be achieved mostly by changing the way experiments are conducted by the students and without additional personnel requirements.

2 Exemplary Experiments

While the first three steps of the Research Data Life Cycle – creating, processing and analysing data – are naturally already included in most of the lab classes, the last three steps – preserving, giving access to and re-using data – are currently missing in most cases. In the following, two exemplary experiments in which a simple modification can enhance the learning outcome regarding DL are described:

2.1 Radioactive decay

An exemplary implementation of all life cycle steps is possible in experiments using radioactive sources, which exist at most universities. Sources like ^{252}Cf with a half-life time of 2.6 a [6] allow the modelling of radioactive decay without additional efforts by applying the different data life cycle steps: Different lab groups perform their experiment in which along the way raw total counts measured from the source are obtained (*creation of data*). The *processing* and *analysis* of this data easily leads to a total count rate. If this rate is collected in digital form from each group (*preserving data*), e.g. in form of a pandas dictionary, following groups can *get access* to this growing data set, and *re-use* it to probe the law of radioactive decay.

2.2 Specific heat of metals

As a second example, in a thermodynamics experiment for the measurement of the temperature dependence of specific heat $c(T)$ of metals, different groups can determine $c(T)$ for one material each (Al, Cu, Pb...) and thereby once experience the whole measurement process on their own. They can then exchange their data (*giving access*) and each analyse the data sets of all groups. By doing so, students on the one hand learn how to record data such that others who did not do the experiment by themselves are able to analyse it, and on the other hand naturally learn how to profit from data which is already taken by others.

The implementation of new topics in existing courses certainly needs to be monitored by appropriate evaluation methods, following the PDCA cycle [7] (plan, do, check, act). It is well-established to offer this task as theses e.g. for graduate or PhD students in education. Continuous evaluation of the lab course allows for immediate reaction to any problem which might arise over time, and is expected to improve the experiment itself, leading to improved student success and a higher rate of students achieving the desired learning outcomes [8]. Additionally, it becomes easier for the teaching staff to adapt to improved technology and methodology in the field.

3 Conclusion

The authors would like to emphasise the strength of the given approach, which is to achieve high impact with only small modifications of existing lab experiments. While the examples given in this contribution are specifically for physics programmes, the underlying idea can be easily adopted by other disciplines: In every programme where students work empirically with data, the corresponding courses can be used as a starting point for teaching RDM and DL competencies by introducing comparably small changes to the curriculum.

To summarise, this contribution specifically outlines how it can succeed to modify common beginners lab experiments in order to add DL competencies to the learning outcomes of the students. Based on the Research Data Life Cycle the authors suggest experiments and outline correspondence to the six stages thereof. For each case, the desired learning outcomes are discussed in detail.

Author contributions

The authors declare that they contributed equally to the present publication throughout all stages of the project. The authors are listed in alphabetical order.

Conflict of interest/Competing interests

The authors declare that they have no competing interests.

Funding

This work has no funding.

Acknowledgements

The authors acknowledge fruitful discussions and collaboration with junge Deutsche Physikalische Gesellschaft (jDPG) and Zusammenkunft aller deutschsprachigen Physikfachschaften (ZaPF).

References

- [1] A. M. Tammaro and V. Casarosa, "Research data management in the curriculum: An interdisciplinary approach," *Procedia Computer Science*, vol. 38, pp. 138–142, 2014, 10th Italian Research Conference on Digital Libraries, IRCDL 2014, ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2014.10.023>.
- [2] R. Schneider, "Research data literacy," in *Worldwide Commonalities and Challenges in Information Literacy Research and Practice*, S. Kurbanoglu, E. Grassian, D. Mizrachi, R. Catts, and S. Špiranec, Eds., Cham: Springer International Publishing, 2013, pp. 134–140, ISBN: 978-3-319-03919-0.
- [3] European Parliament Council, "Recommendation of the European Parliament and of the Council of 23 April 2008 on the establishment of the European qualifications framework for lifelong learning," *Official Journal of the European Union*, vol. 51 C 111, pp. 1–7, 2008.
- [4] I. A. e. a. M. Wilkinson M. Dumontier, "The fair guiding principles for scientific data management and stewardship," *Sci Data*, vol. 3, 160018, 2016. DOI: <https://doi.org/10.1038/sdata.2016.18>.

- [5] J. Bode and P. Jaeger, "Chapter 5 - Redet über die Daten! - Forschungsdatenmanagement und Hochschullehre in der Physik und darüber hinaus," in A. Aljanazrah, S. Brackertz, J. Gehlert, *et al.*, Eds., vol. 1, Oct. 2021, pp. 512–516. DOI: <https://zenodo.org/record/5168524>. [Online]. Available: <https://ojs.dpg-physik.de/index.php/phydid-b/article/view/1180>.
- [6] G. R. CHOPPIN, J.-O. LILJENZIN, and J. RYDBERG, "Chapter 16 - the transuranium elements," in *Radiochemistry and Nuclear Chemistry (Third Edition)*, G. R. CHOPPIN, J.-O. LILJENZIN, and J. RYDBERG, Eds., Third Edition, Woburn: Butterworth-Heinemann, 2002, pp. 415–439, ISBN: 978-0-7506-7463-8. DOI: <https://doi.org/10.1016/B978-075067463-8/50016-9>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780750674638500169>.
- [7] "Deming cycle (pdca)deming cycle plan-do-check act (pdca)(pdca)," in *Encyclopedia of Production and Manufacturing Management*, P. M. Swamidass, Ed. Boston, MA: Springer US, 2000, pp. 155–155, ISBN: 978-1-4020-0612-8. DOI: [10.1007/1-4020-0612-8_229](https://doi.org/10.1007/1-4020-0612-8_229). [Online]. Available: https://doi.org/10.1007/1-4020-0612-8_229.
- [8] Ministerial Conference of the European Higher Education Area (EHEA), "Paris communiqué," 2018. [Online]. Available: https://ehea.info/Upload/document/ministerial_declarations/EHEAParis2018_Communique_final_952771.pdf.

MaRDI

Building Research Data Infrastructures for Mathematics and the Mathematical Sciences

Renita Danabalan¹[\[https://orcid.org/0000-0003-3324-6448\]](https://orcid.org/0000-0003-3324-6448), Michael
Hintermüller¹[\[https://orcid.org/0000-0001-9471-2479\]](https://orcid.org/0000-0001-9471-2479), Thomas Koprucki¹[\[https://orcid.org/0000-0001-6235-9412\]](https://orcid.org/0000-0001-6235-9412),
and Karsten Tabelow¹[\[https://orcid.org/0000-0003-1274-9951\]](https://orcid.org/0000-0003-1274-9951)

¹WIAS Berlin, Germany

Abstract: MaRDI is building a research data infrastructure for mathematics and beyond based on semantic technologies (metadata, ontologies, knowledge graphs) and data repositories. Focusing on the algorithms, models and workflows, the MaRDI infrastructure will connect with other disciplines and NFDI consortia on data processing methods, solving real world problems and support mathematicians on research data management.

Keywords: Research data, mathematics, research data infrastructures, semantic technologies, repositories

1 Introduction

At the heart of many scientific discipline practices lies the processing and analysis of data collected to gain actual scientific insights and/or discoveries. In general, this step can be understood as a sequence of data transformations acting on input creating the output. The output can be used to answer a specific research question and to support research findings. The input and output, together with the data transformations in-between are parts of the factual material necessary to validate research findings, thereby constituting the research data related to specific question.

Next, we illustrate how the concept of data transformations can be applied to research in mathematics. We start with the field of scientific computing that is related to numerical data. For example, solving a linear system of equations $Ax = b$ can be seen as the transformation of the system matrix A and the vector b , via a solver, to the solution vector x . In this case, we can reinterpret the solution process, e.g. the Gaussian elimination, as a transformation of input to output. While the idea of data transformation may be less obvious in other areas of mathematics, it can still be applied; e.g. a computer algebra system, such as Mathematica, ‘transforms’ formulae to formulae. The difference being the data is non-numeric and more complex comprising of both symbolic and exact data.

The diversity of data in mathematics can be attributed to the research process where objects are invented and their relations and their properties are discovered. Data can

range from numerical and tabulated to non-numerical like formulae, symbolic data, models and documents. More importantly, computations with these objects leads to algorithms and research software corresponding to the data transformations as introduced above. The study of their performance is essential to mathematical research and data processing pipelines of the other disciplines where a similar approach to data transformation can be taken; novel findings are acquired by processing and analysing of data.

Although input data might look different, the concept and requirements remain the same. The access to objects, the study of their properties and relations requires standardised data formats, data interoperability and application programming interfaces. With this in mind, the Mathematical Research Data Initiative (MaRDI) will develop a robust Mathematical Research Data Infrastructure that allows for findability and accessibility of objects, investigation of performance of algorithms and a search engine to identify solutions to mathematical problems [1]. This is supposed to support mathematical research fulfilling the needs for data management [2]. Moreover, MaRDI would help to solve real world problems by translating them into mathematical ones. Here, several questions arise, e.g. existence of a mathematical model, availability of solving algorithms, input data or model validity. Bridging the MaRDI infrastructure to disciplines outside of mathematics reduces the amount of time required for finding existing models, algorithms and solvers.

2 The MaRDI layer architecture

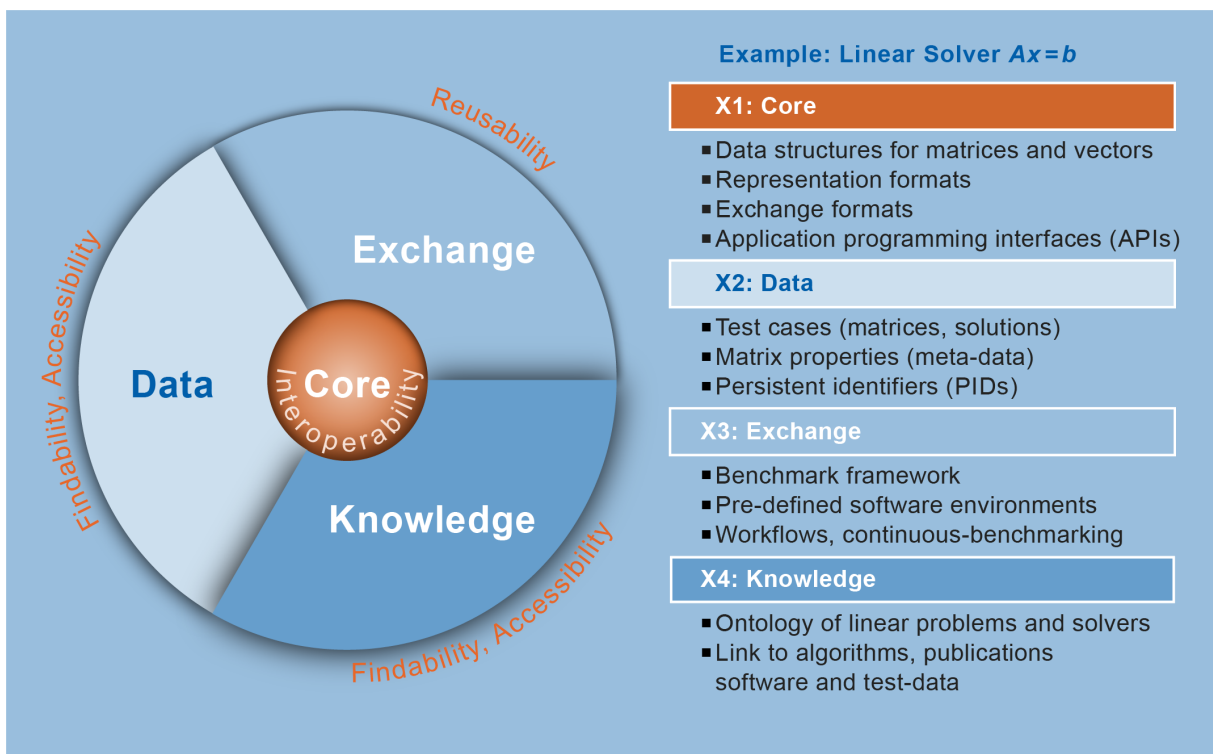


Figure 1. MaRDI layer architecture and FAIR principles. The X1:Core layer enables the interoperability, the X2:Data and X4:Knowledge layer the findability and accessibility and the X3:Exchange layer the reusability.

In order to achieve this goal, our layer architecture (Fig. 1) allows us to define the basic requirements of our infrastructure [1]. This benefits (a) method developers in

running new algorithms on many test problems and also on special collections and (b) users in the generating performance data for selected algorithms or their implementations regarding various test problems from their application context. Additionally, all results of solver runs, using the benchmark framework, can be logged and recorded in the data layer again, e.g. performance data. The collection of these results contributes to systematic evaluation and analysis of the limits of solvers or their implementations.

Explaining the architecture using the example of the linear solver, the first goal of our X1:Core layer (Fig. 1) would be to develop standards for different types of matrices, e.g. full, banded or sparse matrices, which includes in-memory representation formats and file formats for input-output. Second, development of application programming interfaces (APIs) that would implement computations on matrices and vectors and to call the solver. To study the performance of algorithms, the user would run their solver on a set of test cases, consisting of matrices A , vectors b and solutions x . These are provided by X2: Data layer together with metadata schemas for the description of matrix properties (e.g., symmetric positive definite) and the solver (e.g., direct, iterative) as well as query functions for test problems with specific properties (e.g. symmetric, size). An example for such a database is the SuiteSparseCollection [3]. Our X3: Exchange layer, provides an environment that would allow the user to compare the performance, accuracy and efficiency of a solver used for specific matrices. Finally, bringing some structure and relation to the output of the exchange layer, the X4: Knowledge layer uses an ontology of linear problems and solvers to build a knowledge graph for linear problems by linking algorithms to publications, research software and test and performance data. The knowledge graph allows the user to find an appropriate solver for a specific problem or application. Moreover, the MaRDI layer architecture ensures the compliance with the FAIR principles, see Fig. 1.

3 MaRDI task areas

The work in MaRDI is divided into 3 pillars by data categories (Fig. 2): exact and symbolic data (task area (TA) 1:Computer Algebra), floating point data (TA2: Scientific Computing) and data with uncertainties (TA3: Statistics and Machine learning). Our fourth pillar (TA4: Interdisciplinary Mathematics) translates real world problems into mathematical ones through the use of mathematical models solvable by methods from TA1-TA3. Though data types in every TA differ and require individual solutions, the algorithms, mathematical models and workflows are cross-cutting. These cross-cutting topics allow for synergies between the MaRDI task areas and with other disciplines and NFDI consortia. The services, including semantic technologies (metadata, ontologies, knowledge graphs) and data repositories, developed by the task areas will be integrated into the MaRDI portal, our central access point developed by TA5. In parallel, MaRDI's TA6 will engage with and involve the community in the development and use of our services to promote a culture shift towards FAIR science.

Funding

This work was supported by MaRDI, funded by the Deutsche Forschungsgemeinschaft (DFG), project number 460135501, NFDI 29/1 "MaRDI – Mathematische Forschungsdateninitiative".

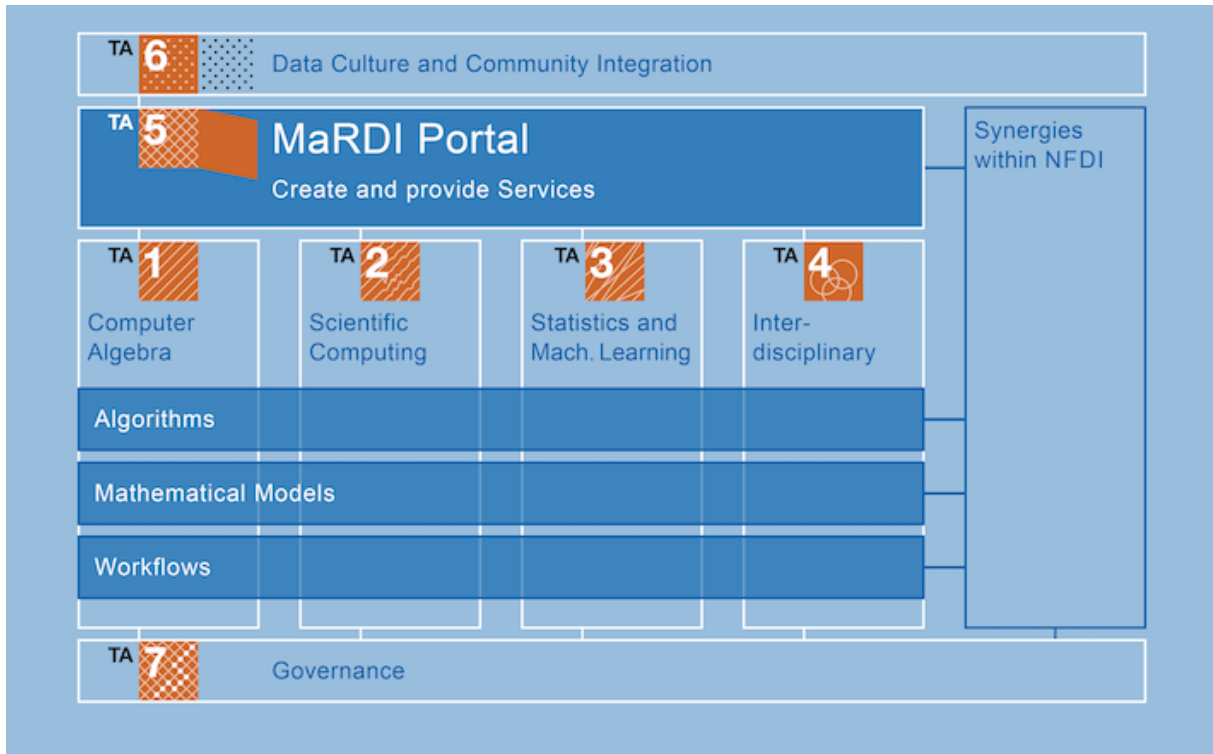


Figure 2. MaRDI organisation. Four task areas address different data categories in mathematics and interdisciplinary research. Task area 5 is dedicated to the MaRDI portal and TA6 deals with the community integration.

Acknowledgements

We thank Thomas Bender for creating the figures.

References

- [1] The MaRDI consortium, *MaRDI: Mathematical Research Data Initiative Proposal*, May 2022. DOI: [10.5281/zenodo.6552436](https://doi.org/10.5281/zenodo.6552436). [Online]. Available: <https://doi.org/10.5281/zenodo.6552436>.
- [2] T. Boege, R. Fritze, C. Görgen, *et al.*, *Research-Data Management Planning in the German Mathematical Community*, 2022. arXiv: [2211.12071](https://arxiv.org/abs/2211.12071) [math.HO].
- [3] T. A. Davis and Y. Hu, "The University of Florida Sparse Matrix Collection," *ACM Trans. Math. Softw.*, vol. 38, no. 1, Dec. 2011, ISSN: 0098-3500. DOI: [10.1145/2049662.2049663](https://doi.org/10.1145/2049662.2049663). [Online]. Available: <https://doi.org/10.1145/2049662.2049663>.

MaRDIFlow: A Workflow Framework for Documentation and Integration of FAIR Computational Experiments

Pavan L. Veluvali¹[\[https://orcid.org/0000-0001-8804-0338\]](https://orcid.org/0000-0001-8804-0338), Jan Heiland¹[\[https://orcid.org/0000-0003-0228-8522\]](https://orcid.org/0000-0003-0228-8522),
and Peter Benner¹[\[https://orcid.org/0000-0003-3362-4103\]](https://orcid.org/0000-0003-3362-4103)

¹Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany.

Abstract: Numerical algorithms and computational tools are essential for managing and analyzing complex data processing tasks. With ever increasing availability of meta-data and parameter-driven simulations, the demand and the need for reliable and automated workflow frameworks to reproduce computational experiments has grown. In this work, we aim to develop a novel computational workflow framework, namely MaRDIFlow, that describes the abstraction of multi-layered workflow components. Herein, we plan to enable and implement scientific computing data FAIRness into actionable guidelines for FAIR computational experiments.

Keywords: FAIR, Computational Workflows, Reproducibility, MaRDIFlow

1 Introduction

Scientific computing has been a cross-disciplinary topic at the borders of applied mathematics, computational sciences and engineering (CSE), as well as other scientific domains involving numerical computations. Likewise, algorithms from numerical mathematics and data are the backbone of simulations in engineering problems, where, practically, different numerical methods are chained to design workflows.

Over the last two decades this has led to the accumulation of significant and frequent difficulties in maintaining solutions and in enabling collaborations within disciplines. We know that the ability to reproduce original research results is contingent on the availability of the original data and methods. As a result, in the past many years, efforts have been devoted towards the development of computational workflows for various scientific applications [1].

In general, a computational workflow is defined as a step-by-step description for accomplishing a scientific objective expressed in terms of tasks and their data dependencies [2]. The complex multi-step methods that are typically used for data collection, data analytics, predictive modeling, and simulation in turn lead to the development of new products. As a research data management (RDM) tool they can also be stored, retrieved for modifications, and subsequently reused in different scenarios with user-defined patterns. Currently, computational workflows offer graphical interfaces with high-level mechanisms for composition as well as traditional text-based programming interfaces.

However, a key challenge for computational scientists is building a framework for creating, maintaining, and accessing reusable workflows [3]. While tools and programming interfaces are important aspects in computational science and engineering, integrating them into a workflow framework can further express details about meta data and task dependencies, respectively. Nonetheless, motivated by demands of the mathematical community and other disciplines, MaRDI (Mathematical Research Data Initiative) [4], the consortial initiative of mathematical sciences, aims to set standards for the design of confirmable workflows.

In this regard, as a part of the MaRDI consortium on research data management in mathematical sciences [4], we present a novel computational framework, namely MaRDIFlow, that focuses on automation of abstracting meta-data embedded in an ontology of mathematical objects while negating the underlying execution and environment dependencies into multi-layered descriptions.

2 MaRDIFlow

The overall objective of our workflow framework is to provide a programming environment that simplifies the effort required by users or scientists to orchestrate a FAIR computational experiment. In MaRDIFlow, the workflow components are considered as abstract objects which are in turn described by their input to output behavior and as well as by their corresponding metadata. Through metadata and by matching the I/O interfaces, the objects are chained together to form a computational workflow. Herein, input stands for the parameter that sets up the current part of the workflow, and output denotes the final/intermediate result which is passed on to the next component. In this way, a workflow component can be described in a multi-layered fashion, namely via a mathematical/physical model, via a model that has been inferred from data, or via plain data, see Figure. 1.

For a systematic description of CSE components, we plan to incorporate models of different kind, code, and data equivalently and redundantly. One of the important benefits of combining data and code lies in the flexible treatment of the associated simulation data. For example, the storage requirements of huge time series can be reduced by replacing the full data by parts and associated code that can provide the missing points on demand. Also, simulation parts that are defined as the result of empirical statistics can be provisioned with the relevant code and statistical information and further improved as needed. Another benefit of the input/output perspective is the interchangeability of the concrete realization so that, e.g., for reproduction, a closed-source implementation can be substituted by an open-source equivalent.

In order to present the description of individual workflow components, we produce show cases with different meta-data for redundant and reproducible mathematical models. As a minimum working example we incorporate a two-dimensional model for Cahn-Hilliard equation [5] that simulates the phase-separation of a binary A-B alloy. Moreover, the developed computational workflow framework adheres to FAIR principles [6], such that abstracted components are Findable, Accessible, Interoperable, and Reusable. In addition, going forward we plan to provide our RDM tool through electronic lab notebooks (ELNs) with minimal adjustments and user-friendly guidelines.

Lastly, we believe that the CSE workflow description presented here as a part of the MaRDI consortium serves a scientific tool for research data management in numerical mathematics.

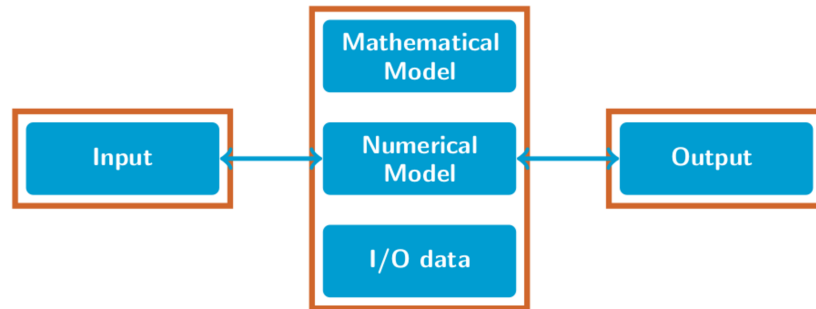


Figure 1. MaRDIFlow: A CSE workflow framework for documentation and integration of FAIR computational experiments

Data availability statement

Results presented in this work are apart of an ongoing investigation, however a working prototype of our workflow framework is available and documented at <https://zenodo.org/record/78635>

Competing interests

The authors declare that they have no competing interests.

Funding

Authors are supported by MaRDI, funded by the Deutsche Forschungsgemeinschaft (DFG), project number 460135501, NFDI 29/1 “MaRDI – Mathematische Forschungsdateninitiative”

References

- [1] D. Talia, “Workflow systems for science: Concepts and tools,” *International Scholarly Research Notices*, vol. 2013, 2013.
- [2] C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, *et al.*, “FAIR Computational Workflows,” *Data Intelligence*, vol. 2, no. 1-2, pp. 108–121, 2020.
- [3] M. Wolf, J. Logan, K. Mehta, *et al.*, “Reusability first: Toward FAIR workflows,” in *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, IEEE, 2021, pp. 444–455.
- [4] MaRDI. “Mathematic research data initiative.” (2021), [Online]. Available: <https://www.mardi4nfdi.de>.
- [5] J. W. Cahn and J. E. Hilliard, “Free energy of a nonuniform system. I: Interfacial free energy,” *The Journal of chemical physics*, vol. 28, no. 2, pp. 258–267, 1958.
- [6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, “The FAIR guiding principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

Overarching Data Management Ecosystem at HZDR

From Small Experiments to Large-Scale Research Facilities

Oliver Knodel¹[\[https://orcid.org/0000-0001-8174-7795\]](https://orcid.org/0000-0001-8174-7795), Thomas Gruber¹[\[https://orcid.org/0000-0001-6940-2065\]](https://orcid.org/0000-0001-6940-2065),
Jeffrey Kelling¹[\[https://orcid.org/0000-0003-1761-2591\]](https://orcid.org/0000-0003-1761-2591), Mani Lokamani¹[\[https://orcid.org/0000-0001-8679-5905\]](https://orcid.org/0000-0001-8679-5905),
Stefan Müller¹[\[https://orcid.org/0000-0001-6273-7102\]](https://orcid.org/0000-0001-6273-7102), David Pape¹[\[https://orcid.org/0000-0002-3145-9880\]](https://orcid.org/0000-0002-3145-9880), Martin
Voigt^{1,2}[\[https://orcid.org/0000-0001-5556-838X\]](https://orcid.org/0000-0001-5556-838X), and Guido Juckeland¹[\[https://orcid.org/0000-0002-9935-4428\]](https://orcid.org/0000-0002-9935-4428)

¹Helmholtz-Zentrum Dresden – Rossendorf, Dresden, Germany

²Technische Universität Dresden, Germany

Abstract: When dealing with research data management, researchers at Helmholtz-Zentrum Dresden – Rossendorf (HZDR) face a variety of systems and tools. These range from the project planning phase (proposal management, data management plans and policies), over documentation during the experiment or simulation campaign, to the publication (collaborative authoring tools, metadata catalogs, publication systems, data repositories). In addition, modern research projects usually are required to interact with a variety of software stacks and workflow management systems to allow comprehensible and FAIR science on the underlying IT infrastructure (HPC, data storage, network file systems, archival). This article first demonstrates the data management systems and services provided at HZDR, followed by an overview of a self-developed guidance system. It is concluded by a real-world example.

Keywords: research data management, data life cycle, workflows, metadata, FAIR, data provenance, HELIPORT

1 Data Management Ecosystem at HZDR

Over the last years, we have been developing a uniform data management ecosystem aiming to make the entire life cycle of a scientific project – from the submission of a beamtime proposal to the publication of produced datasets – comprehensible according to the FAIR principles [1]. Figure 1 shows this ecosystem consisting of the various systems and services in our IT landscape. The systems access the underlying hardware in the data centre without abstraction via the common interfaces.

Visiting scientists typically start their journey through our infrastructure by submitting a proposal to the proposal management system *Gate* [3] on the left side of Figure 1. After acceptance, the first basic set of metadata, such as researcher names and Open Researcher and Contributor IDs (ORCID) [4], title and abstract of the experiment, as well as experiment type or the facility used, are known. Using this initial data, a data management plan (DMP) can be generated with the help of the Research Data Management Organiser (RDMO) [5].

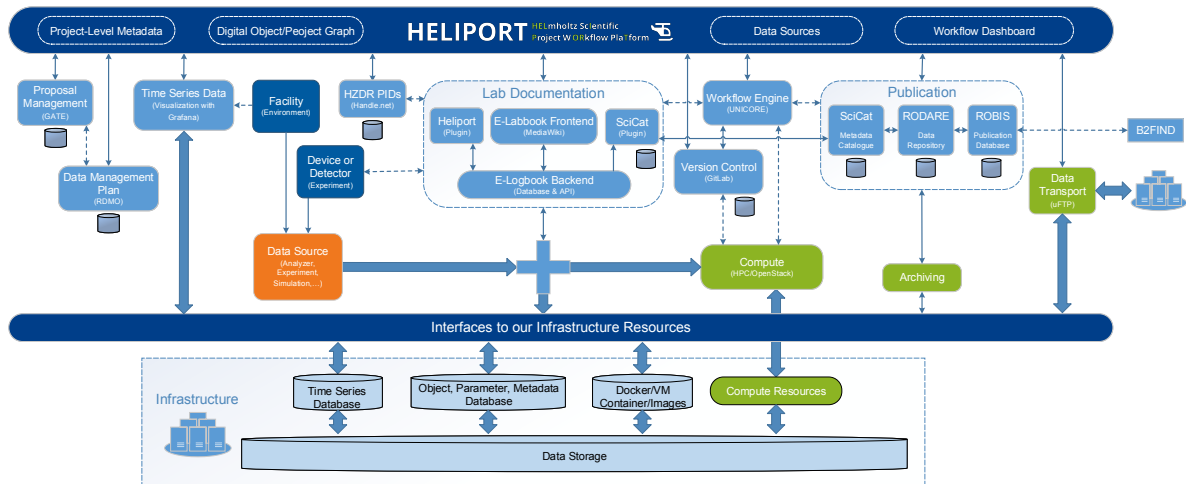


Figure 1. Top-Level Architecture of the HZDR Data Management ecosystem with the various underlying systems and services. [2]

Experiment control, data acquisition and facility (meta)data are provided by different (experiment-specific) subsystems, and are typically stored in a time series database. For visualization, a central Grafana OSS [6] instance is available. Together with automated and user-defined documentation in our e-labbook (based on Semantic MediaWiki [7]), the actual data from an experiment or a simulation (digital twin), is used for pre- and post-analysis workflows on a high performance computing (HPC) cluster. The access to our HPC cluster is provided by the Uniform Interface to Computing Resources (UNICORE) [8]. A combination of metadata- and data-repository is used for the publication of results and the registration in our publication database Rossendorf Bibliography System (Robis). The experiment-specific metadata can be transferred or entered directly into a metadata catalog based on SciCat [9] with a direct link to the data available in our data repository Rossendorf Data Repository (Rodare) [10], which is based on Zenodo [11] and Invenio [12].

We plan to provide additional Digital Object Identifiers (DOIs) for our large scale facilities and beamlines. In the past, beamline scientists created descriptions of their facilities in the Journal of large-scale research facilities (JLSRF) as for instance [13]. Due to the fact that most of our facilities have not published a citable description, uniform landing pages are created, such as [14] for the NELBE facility. This information, as well as the proposal metadata, can be attached to the publication of the dataset on Rodare as a related identifier and the environment of an experiment where a dataset was created can be comprehended and reproduced.

2 Guidance System – HELIPORT

Over the last years, the overarching layer (or guidance system) of our data management ecosystem, introduced in section 1, received the name *HELIPORT*. This abstraction layer HELIPORT [15], [16] is an overall data management solution that aims at making the steps of the entire research experiment's life-cycle findable, accessible, interoperable and reusable according to the FAIR principles. In doing so, it makes the components involved in the project discoverable for new team members and provides valuable functionality to exchange data and metadata between systems.

Among other information, HELIPOINT integrates documentation, scientific workflows and their results, and the final publication of the primary data and research results – all via already established solutions. Integration is accomplished by presenting the researchers with a high-level overview to keep all aspects of the experiment in mind.

Computational agents can interact with HELIPOINT via a REST API that allows access to all components. Furthermore, all aspects of the experiment are registered as digital objects with landing pages that contain metadata in various standardized formats and schemas. Thus, the metadata is readable for both humans and machines. An overall digital object graph, combining the metadata harvested from all sources, provides the scientists with a visual representation of interactions and relations between their digital objects. Additionally, the project timeline lists all digital objects ordered by the time they were created. Through integrated computational workflow systems, HELIPOINT can automate calculations using the collected metadata. We also created a HELIPOINT team on WorkflowHub [17] to allow users to exchange their workflows.

By visualizing all aspects of large-scale research experiments, HELIPOINT enables deeper insights into a comprehensible data provenance with the chance of raising awareness for data management.

3 Data Management View of the TELBE Experiment in HELIPOINT

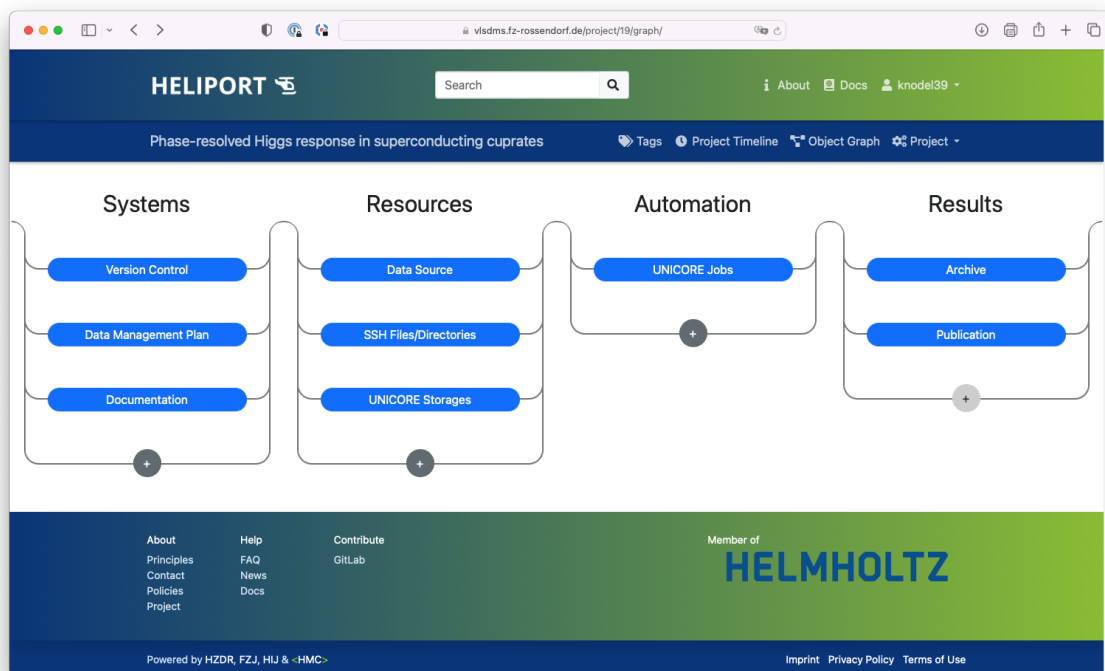


Figure 2. Overview page with all services and systems used in an exemplary High-Field High-Repetition-Rate Terahertz facility (TELBE) project.

As a first example to demonstrate the benefits of HELIPOINT (see section 2) we used an experiment from the High-Field High-Repetition-Rate Terahertz facility (TELBE) [18], which is part of ELBE [13]. The HELIPOINT project overview for this experiment is shown in Figure 2. A variety of the systems mentioned in section 1 are present here, e. g.:

Version Control gathers the source code for post-processing scripts stored in Helmholtz Codebase (a GitLab for the entire Helmholtz association).

Documentation: The TELBE group documents their experiment and record scientific metadata within the e-labbook. It is linked here to be easily findable and updateable.

UNICORE Jobs: This section provides access to the output logs of the UNICORE jobs that run post-processing, as well as some management functionality for these jobs implemented through the UNICORE API.

Publication: Here, the data publications of the post-processed primary data can be found. Datasets registered under **Resources** can be tagged and automatically published on Rodare.

This project also makes use of the HELIPOINT REST API: The experiment control is implemented in LabView and makes API requests to HELIPOINT for current experiment metadata. The metadata is then used to start new post-processing runs on the HPC cluster via UNICORE after each measurement. A callback URL provided by HELIPOINT even allows registration of new jobs and datasets with the project without any user intervention. UNICORE posts status updates about running jobs to this URL and HELIPOINT reacts accordingly.

In cooperation with the TELBE experiment we created a proof-of-concept to demonstrate the potential of our guidance system HELIPOINT to abstract from the complex data management ecosystem it builds upon.

Author contributions

The authors contributed equally to this work.

Competing interests

The authors declare that they have no competing interests.

Funding

Our overall data management strategy and our ecosystem was funded by HZDR and the Helmholtz Association.

The HELIPOINT project (ZT-I-PF-3-021) was funded by the "Initiating and Networking Fund of the Helmholtz Association" in the framework of the first "Helmholtz Metadata Collaboration" project call 2021.

Acknowledgements

We especially thank Uwe Konrad, Maik Fiedler, Jürgen Grzondziel and Jan-Christoph Deinert from HZDR in supporting our initiative, as well as the HMC colleagues responsible for the hub *Matter* for the enriching discussions. Last but not least, we would like to thank the entire HELIPOINT team from Forschungszentrum Jülich, Helmholtz Institute Jena and HZDR.

References

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "Comment: The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, pp. 1–9, 2016, ISSN: 20524463. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
- [2] O. Knodel, T. Gruber, J. Kelling, *et al.*, *HZDR Data Management Strategy — Top-Level Architecture*, Feb. 2023. DOI: [10.14278/rodare.2162](https://doi.org/10.14278/rodare.2162). [Online]. Available: <https://doi.org/10.14278/rodare.2162>.
- [3] User Office, Helmholtz-Zentrum Dresden-Rossendorf, *HZDR Proposal Management System*. [Online]. Available: <https://gate.hzdr.de/user/>.
- [4] *Open Researcher and Contributor ID (ORCID)*. [Online]. Available: <https://orcid.org>.
- [5] J. Klar, O. Michaelis, C. Engelhardt, *et al.*, *Research Data Management Organizer (RDMO)*, version 1.3, Oct. 2020. DOI: [10.5281/zenodo.596581](https://doi.org/10.5281/zenodo.596581). [Online]. Available: <https://github.com/rdmorganiser/rdmo>.
- [6] Grafana Labs, *Grafana OSS | Metrics, logs, traces, and more*. [Online]. Available: <https://grafana.com/oss/grafana/>.
- [7] M. Krötzsch, D. Vrandečić, and M. Völkel, "Semantic mediawiki," in *International semantic web conference*, Springer, 2006, pp. 935–942. DOI: https://doi.org/10.1007/978-3-642-19797-0_16.
- [8] K. Benedyczak, B. Schuller, M. Petrova-El Sayed, J. Rybicki, and R. Grunzke, "Unicore 7—middleware services for distributed and federated computing," in *2016 International Conference on High Performance Computing & Simulation (HPCS)*, IEEE, 2016, pp. 613–620. [Online]. Available: <http://hdl.handle.net/2128/12214>.
- [9] *Scicat metadata catalogue*. [Online]. Available: <https://scicatproject.github.io>.
- [10] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), *RODARE - Rossendorf Data Repository*. DOI: <http://doi.org/10.17616/R3BR40>.
- [11] European Organization For Nuclear Research and OpenAIRE, *Zenodo*, en, 2013. DOI: [10.25495/7GXK-RD71](https://doi.org/10.25495/7GXK-RD71). [Online]. Available: <https://www.zenodo.org/>.
- [12] *Invenio - Powering Open Science*. [Online]. Available: <https://inveniosoftware.org>.
- [13] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), *ELBE Center for High-Power Radiation Sources*. DOI: <https://doi.org/10.17815/jlsrf-2-58>.
- [14] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), *Description of the NELBE facility*. DOI: <https://doi.org/10.58065/24017>.
- [15] O. Knodel, M. Voigt, R. Ufer, *et al.*, "Heliport: A portable platform for FAIR Workflow | Metadata | Scientific Project Lifecycle management and everything," in *Proceedings of the 4th International Workshop on Practical Reproducible Evaluation of Computer Systems*, ser. P-RECS '21, Virtual Event, Sweden: Association for Computing Machinery, 2021, pp. 9–14, ISBN: 9781450383950. DOI: [10.1145/3456287.3465477](https://doi.org/10.1145/3456287.3465477). [Online]. Available: <https://doi.org/10.1145/3456287.3465477>.
- [16] M. Voigt, R. Ufer, W. Schacht, *et al.*, *HELIPORT (HELMholtz Scientific Project WORKflow PlaTform)*, Nov. 2022. DOI: [10.14278/rodare.1970](https://doi.org/10.14278/rodare.1970). [Online]. Available: <https://doi.org/10.14278/rodare.1970>.
- [17] University of Manchester and HITS gGmbH, *HELIPORT team on workflowhub.eu*. [Online]. Available: <https://workflowhub.eu/projects/156>.
- [18] M. Helm, S. Winnerl, A. Pashkin, *et al.*, "The elbe infrared and thz facility at helmholtz-zentrum dresden-rossendorf," *The European Physical Journal Plus*, vol. 138, no. 2, p. 158, 2023. DOI: <https://doi.org/10.1140/epjp/s13360-023-03720-z>.

Two-Step Approach in Metadata Management for Data Publications at Research Centres

Thomas Gruber¹[\[https://orcid.org/0000-0001-6940-2065\]](https://orcid.org/0000-0001-6940-2065), Hans-Peter Schlenvoigt¹[\[https://orcid.org/0000-0003-4400-1315\]](https://orcid.org/0000-0003-4400-1315), Oliver Knodel¹[\[https://orcid.org/0000-0001-8174-7795\]](https://orcid.org/0000-0001-8174-7795), Kristin Tippey¹[\[https://orcid.org/0000-0002-9261-7643\]](https://orcid.org/0000-0002-9261-7643), and Guido Juckeland¹[\[https://orcid.org/0000-0002-9935-4428\]](https://orcid.org/0000-0002-9935-4428)

¹Helmholtz-Zentrum Dresden – Rossendorf, Dresden, Germany

Abstract: Data repositories like Zenodo have a limited list of metadata to search for. Metadata catalogues are designed to provide a community-specific parameters search, but their deployment has just started. These catalogues require metadata standards for interoperability, which in turn are often in development for many communities. To support publications with a metadata standard in the future, a two-step concept is presented in this article. It discusses how the electronic documentation should be constructed, in order to convert this later into a standardised schema for publication. We will present examples from the laser-plasma community for both steps, firstly how we deal with the complex challenges of metadata management and secondly for methods for developing metadata schemas.

Keywords: Data Management, Workflows, Metadata, Data Provenance

1 General Concept of metadata handling

Data publications via repositories like Zenodo [1] are becoming increasingly popular. Many centres around the world are setting up their own instances. Within these repositories one can search for metadata according to DataCite [2]. However, queries on domain-specific metadata require other repositories like metadata catalogues. A key challenge in their development is to deal with the different, domain-specific metadata schemata, not only for metadata aggregation, but also to construct a usable search engine.

The repositories need to be generic in the usage of metadata schemas since those often do not yet exist or are in development, and are likely to evolve later on as knowledge and techniques advance. On the other hand, scientists want to document what they have now – waiting for an established metadata schema is no choice! Therefore we propose an iterative approach. First we collect all metadata currently known in a structured manner. The second step, for data publication, would be restructuring the metadata into a new, documented schema, such that a third party can use the metadata. This comes with several advantages. The documentation of experiments and simulations remains independent from any schema development. Later reformatting would be possible if the schema changes. What it needs is a mapping between local

structured metadata and public metadata schema. A curation and selection process can be implemented between data taking and publication. This comes at the cost of additional work of transformation and managing the mapping configuration.

The diversity of experiments at a research centre demands a high flexibility on the electronic documentation of experiments and simulations. Usually, the starting point is an electronic lab notebook (ELN), but additional databases can be involved to provide full documentation. All sources of electronic documentations will further be called electronic lab documentation and include the interconnections via IDs or links, which brings all metadata and data together. Each element of the e-lab documentation can store and provide metadata and data in a structured fashion like key-value pairs. An ELN is basically a database with a user friendly web front end to manually access and enter metadata. The ELN comes along with special features, which are the reason for the wide variety of ELNs.

At the Helmholtz-Zentrum Dresden – Rossendorf (HZDR), we focus on Mediawiki (based on Semantic MediaWiki [3]) as an ELN and integrate it into many systems which generate or accept metadata (Figure 1). Page templates allow to define structured data types inside the pages as key-value pairs. Input forms can be specifically designed to help and guide the scientist for manual entry of structured metadata. The web front end provides also a html editor to enter unstructured (meta)data, figures, screenshots etc. It is also possible to include or link information from external sources like databases, depending on the specific lab. If an experiment is performed by a device control software like Labview, that can be deployed to automatically send all data and metadata to the e-lab documentation. Likewise, computer simulations can be amended similarly to automatically submit the input parameters and metadata to the e-lab documentation.

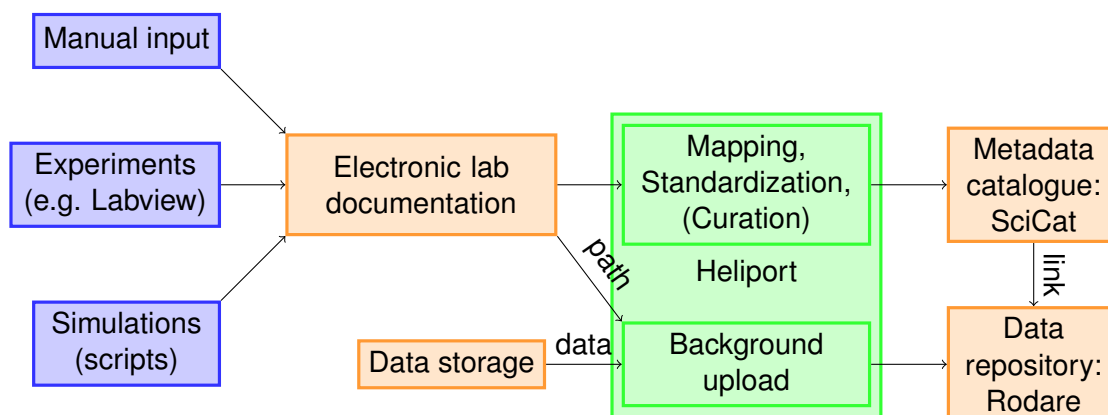


Figure 1. Metadata flow at HZDR. Data and metadata sources (blue) send the (meta)data to the storages (orange), initially to those on the left side. The workflow from primary to publicly available storage is depicted in green.

Once the information is collected in a structured manner, database-like queries can be used to retrieve specific information, not only within the e-lab documentation but also through the API. This is essential to forward the stored information to other services like the metadata catalogue SciCat [4]. A publication workflow could extract the metadata of the selected datasets, transform it into a SciCat [4] publication including all the metadata of the linked sample, project and instrument, by using the mapping configuration to create a certain metadata schema. In parallel the linked data storage path in the e-lab documentation is used to upload the data from the storage to our Rossendorf Data Repository (Rodare) [5], which is based on Zenodo [1] and Invenio [6]. Within this step the data could be combined with the metadata to create a HDF5 or Nexus file

and upload this instead. The data is then referenced within SciCat as a direct link. To manage the workflows and documentation, HELIPOINT [7] is used at HZDR.

In the following, we highlight two example activities of metadata management at HZDR where we touch different communities.

2 Example for (meta)data aggregation: DAPHNE4NFDI

The [Data from PHoton and Neutron Experiments for NFDI \(DAPHNE4NFDI\)](#) project is centred at large-scale research facilities with photons and neutrons and aims for improved metadata management, commonalities in repositories and databases as well as establishing a software ecosystem for analysis software. HZDR's high-intensity laser group participates in DAPHNE4NFDI due to their research of laser-driven plasmas with XFELs. That research of laser-plasma interaction always employs both experiments and numerical simulations. Only the latter allow for insights into the micro-physics (on nanometer spatial and femtosecond temporal scales) whereas the former includes all effects without assumptions, approximations or models.

Within DAPHNE4NFDI, HZDR is working on improving metadata generation and capturing for both simulations and experiments. For simulations, input parameters, based on the Particle-In-Cell Modeling Interface (PICMI) [8], are filed as metadata into a database where simulation output is the data. Thereby, a searchable collection of already conducted simulations can be generated.

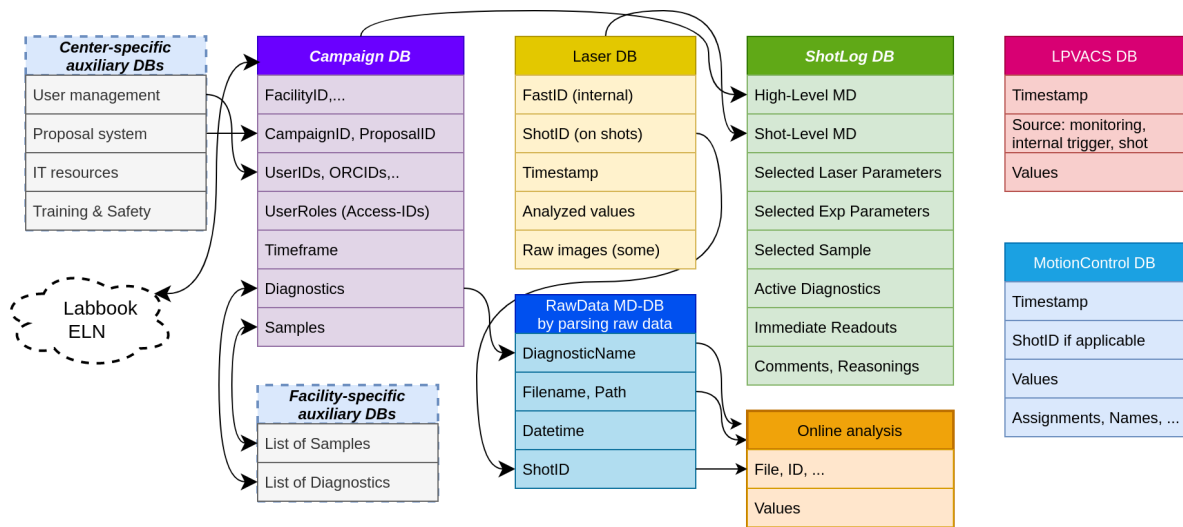


Figure 2. Current concept of a database system specific to laser-particle acceleration experiments at HZDR. This layout provides a concrete example of aggregating all currently available data and metadata as an electronic lab documentation as the first step in our approach.

For experiments, a database system for data and metadata is planned, see Figure 2. So far, raw data - mostly images - is initially stored on local file systems and subsequently aggregated into a file repository. There, structural metadata is encoded in the path structure but not independently searchable. In a first step, the path structure can be parsed into a database (dark blue). In parallel exists a database of the drive laser parameters (yellow). The sequence of laser shots is manually entered into the Shot-Log database (green), alongside with comments on observations and decisions. This database is the cornerstone for later analysis. Further metadata and data sources,

e.g. campaign IDs for metadata or instrument configuration for data, will be added incrementally.

3 Example for metadata schema development: HELPMI

The [HElMholtz Laser Plasma Metadata Initiative \(HELPMI\)](#) project is enabled by the "[Helmholtz Metadata Collaboration](#)" and aims at developing a metadata schema for laser-plasma experimental data, starting from the existing data and metadata standards openPMD [9] and NeXus [10]. So far, openPMD is widely used for laser-plasma simulation data, while NeXus is for experimental data from photon and neutron facilities. However, there is no data and metadata standard for experimental data from laser-plasma research.

HELPMI will develop - together with the global community - a glossary for laser-plasma experimental data. This is an important step towards data publications following the F.A.I.R. principles [11], such that there are commonly accepted definitions of terms in conjunction with domain-specific hierarchies.

HELPMI will make openPMD substantially extensible for custom hierarchies and will furthermore adopt the NeXus format, in particular in regard of geometry description and raw data and processed data handling. In combination with the glossary, openPMD can become a metadata standard for laser-plasma experiments and simulations. Furthermore there is potential that NeXus and openPMD - standards of two different scientific communities - can become interoperable.

Ultimately, once a metadata standard exists, the aggregated data and metadata can be mapped to that standard for F.A.I.R. data publications.

Author contributions

These authors contributed equally to this work.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the consortium DAPHNE4NFDI in the context of the work of the NFDI e.V. The consortium is funded by the DFG - project number 460248799. The HELPMI project (ZT-I-PF-3-066) was funded by the "Initiative and Networking Fund" of the Helmholtz Association in the framework of the "Helmholtz Metadata Collaboration" project call 2022.

References

- [1] European Organization For Nuclear Research and OpenAIRE, *Zenodo*, en, 2013. DOI: [10.25495/7GXK-RD71](https://doi.org/10.25495/7GXK-RD71). [Online]. Available: <https://www.zenodo.org/>.
- [2] D. M. W. Group *et al.*, "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs," 2021. DOI: <https://doi.org/10.14454/3w3z-sa82>.

- [3] M. Krötzsch, D. Vrandečić, and M. Völkel, "Semantic mediawiki," in *International semantic web conference*, Springer, 2006, pp. 935–942. DOI: https://doi.org/10.1007/978-3-642-19797-0_16.
- [4] *SciCat Metadata Catalogue*. [Online]. Available: <https://scicatproject.github.io>.
- [5] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), *RODARE - Rossendorf Data Repository*. DOI: <http://doi.org/10.17616/R3BR40>.
- [6] *Invenio - Powering Open Science*. [Online]. Available: <https://inveniosoftware.org>.
- [7] O. Knodel, M. Voigt, R. Ufer, et al., "Heliport: A portable platform for FAIR Workflow — Metadata — Scientific Project Lifecycle management and everything," in *Proceedings of the 4th International Workshop on Practical Reproducible Evaluation of Computer Systems*, ser. P-RECS '21, Virtual Event, Sweden: Association for Computing Machinery, 2021, pp. 9–14, ISBN: 9781450383950. DOI: [10.1145/3456287.3465477](https://doi.org/10.1145/3456287.3465477). [Online]. Available: <https://doi.org/10.1145/3456287.3465477>.
- [8] *Particle-In-Cell Modeling Interface*. [Online]. Available: <https://picmi-standard.github.io/>.
- [9] *openPMD: A meta-data standard*. [Online]. Available: <https://www.openpmd.org/>.
- [10] *the NeXus Data Format*. [Online]. Available: <https://www.nexusformat.org/>.
- [11] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al., "Comment: The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, pp. 1–9, 2016, ISSN: 20524463. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).

Harmonising, Harvesting, and Searching Metadata across a Repository Federation

Steffen Neumann¹[\[https://orcid.org/0000-0002-7899-7192\]](https://orcid.org/0000-0002-7899-7192), Felix Bach²[\[https://orcid.org/0000-0002-5035-7978\]](https://orcid.org/0000-0002-5035-7978),
Leyla Castro³[\[https://orcid.org/0000-0003-3986-0510\]](https://orcid.org/0000-0003-3986-0510), Tillmann G. Fischer¹[\[https://orcid.org/0000-0003-4480-8661\]](https://orcid.org/0000-0003-4480-8661),
Stefan Hofmann²[\[https://orcid.org/0000-0003-0790-112X\]](https://orcid.org/0000-0003-0790-112X), Pei-Chi Huang⁴[\[https://orcid.org/0000-0002-9976-4507\]](https://orcid.org/0000-0002-9976-4507),
Nicole Jung⁴[\[https://orcid.org/0000-0001-9513-2468\]](https://orcid.org/0000-0001-9513-2468), Bhavin Katabathuni⁶[\[https://orcid.org/0009-0003-1198-9969\]](https://orcid.org/0009-0003-1198-9969),
Fabian Mauz¹[\[https://orcid.org/0000-0003-4673-5494\]](https://orcid.org/0000-0003-4673-5494), Rene Meier¹[\[https://orcid.org/0000-0002-1501-1349\]](https://orcid.org/0000-0002-1501-1349),
Venkata Chandrasekhar Nainala⁵[\[https://orcid.org/0000-0002-2564-3243\]](https://orcid.org/0000-0002-2564-3243),
Noura Rayya⁵[\[https://orcid.org/0009-0001-5998-5030\]](https://orcid.org/0009-0001-5998-5030), Christoph Steinbeck⁵[\[https://orcid.org/0000-0001-6966-0814\]](https://orcid.org/0000-0001-6966-0814),
and Oliver Koehler⁶[\[https://orcid.org/0000-0003-3385-4232\]](https://orcid.org/0000-0003-3385-4232)

¹ Leibniz Institute of Plant Biochemistry, Halle, Germany, <https://ror.org/01mzk5576>

² FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Karlsruhe, Germany, <https://ror.org/04z92tv25>

³ ZB Med Information Centre for Life Sciences, Cologne, Germany, <https://ror.org/0259fwx54>

⁴ Karlsruhe Institute of Technology, Karlsruhe, Germany, <https://ror.org/04t3en479>

⁵ Friedrich-Schiller-University, Jena, Germany, <https://ror.org/05qpz1x62>

⁶ TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany, <https://ror.org/04aj4c181>

Abstract. The collection of metadata for research data is an important aspect in the FAIR principles. The schema.org and Bioschemas initiatives created a vocabulary to embed markup for many different types, including BioChemEntity, ChemicalSubstance, Gene, MolecularEntity, Protein, and others relevant in the Natural and Life Sciences with immediate benefits for findability of data packages. To bridge the gap between the worlds of semantic-web-driven JSON+LD metadata on the one hand, and established but separately developed interface services in libraries, we have designed an architecture for harmonising, federating and harvesting metadata from several resources. Our approach is to serve JSON+LD embedded in an XML container through a central OAI-Provider. Several resources in NFDI4Chem provide such domain-specific metadata. The CKAN-based NFDI4Chem search service can harvest this metadata using an OAI-PMH harvester extension that can extract the XML-encapsulated JSON+LD metadata, and has search capabilities relevant in the chemistry domain. We invite the community to collaborate and reach a critical mass of providers and consumers in the NFDI.

Keywords: Metadata, Structured Markup, JSON+LD, schema.org, Bioschemas, OAI-PMH, Harvesting

1. Background

Research data is a critical component of scientific inquiry, and its value lies in its ability to support the reproducibility, transparency, and reuse of research findings. By sharing research data, researchers enable others to verify their findings, build upon them, and contribute to the development of new knowledge. Several generic and domain-specific data repositories have been introduced over the years to publish data. DataCite not only provides persistent identifiers

(DOIs) for research datasets and data publications, but also the DataCite Metadata Schema as a standard format for a generic and discipline-independent description of research data. This enables easy discovery, access, and reuse. By adopting the DataCite Metadata Schema, researchers can ensure their data is properly documented and discoverable, which increases the visibility and impact of their research. Nevertheless, DataCite Schema is somewhat limited when it comes to expressing domain-specific metadata like molecular entities, chemical structure information, analytical methods, or processes.

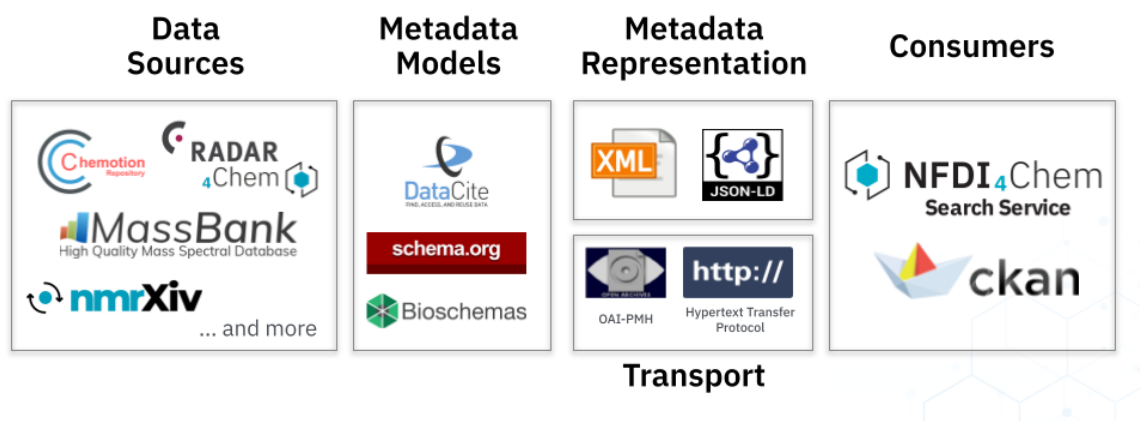
The schema.org initiative created a vocabulary that allows website owners to embed markup for many different types like people, events, organisations, or places using embedded JSON+LD, with the immediate benefit of better findability [1]. Bioschemas [2] is a community effort to improve the FAIRness of data of resources in the Life sciences by defining specific metadata schemas as JSON+LD and exposing that metadata from web based resources that have adopted it. To this end, it offers some tailored types that are readily applicable in many disciplines in the natural and life-sciences. Some of the types (e.g., BioChemEntity, ChemicalSubstance, Gene, MolecularEntity, Protein, and Taxon) have been picked up into schema.org. In addition to the types, Bioschemas also offers some validation and harvesting tools, making it easier to comply with specifications and consume the markup.

Libraries have been sharing the metadata about their own content to improve the findability of a book or article across libraries. The de facto standard to support discovery, presentation, and analysis of data originating from compliant archives is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [3].

Within the NFDI, the NFDI4Chem consortium has set out to create an open and FAIR infrastructure for research data management in chemistry, initially focusing on data related to molecules and reactions including data for their experimental and theoretical characterisation [4]. To successfully achieve this goal, we need to merge the two aforementioned technologies and consolidate the metadata into a central location where it can be indexed by chemistry aware search services, and equally well by generic data discovery services.

2. Modeling of chemistry-specific metadata and Design of metadata workflows

The NFDI4Chem federation of data repositories has initially started with existing repositories that had already implemented metadata formats and interfaces. These repositories were the first ones considered for harvesting and aggregating metadata into a harmonised metadata store and NFDI4Chem search service.



To demonstrate the feasibility, but also challenges and requirements when designing an architecture and large-scale system to provide harmonised, federated, disparate metadata from different realms and allow to harvest, search, and integrate this information, we have connected several existing methods and systems for chemistry research data.

The NFDI4Chem has several resources which serve (or plan to serve) Schema.org, i.e., Bioschemas metadata. MassBank [5] is an open spectral database, and it has been supporting Bioschemas markup for the individual records as Dataset and MolecularEntity since 2018. The Chemotion repository is well integrated with the Chemotion Electronic Lab Notebook. Since the repository already has a powerful REST API, we were able to develop a light-weight conversion from the Chemotion data types to schema.org MolecularEntity and Dataset types.

Designed for NMR data, nmrXiv is an open repository for FAIR NMR spectroscopy data. The data model is closely following the ISA model [6]. Bioschemas markup was integrated early in the architecture of nmrXiv. RADAR4Chem is a generic repository that can capture domain-specific metadata. The following table summarises supported types.

	Chemotion Repository	nmrXiv	MassBank	RADAR4Chem
DataCatalog	✓	✓	✓	
Study	✓	✓		
CreativeWork	✓	✓		
Person	✓	✓		✓
LabProtocol	✓			
ChemicalSubstance	✓	✓	✓	
MolecularEntity	✓	✓		
Dataset	✓	✓	✓	✓
DataDownload	✓	✓		

To bridge the gap between the semantic-web driven JSON+LD metadata and the established workflows in libraries, we need to design an approach to serve JSON+LD via OAI-PMH. Since, by design, the OAI responses are XML documents, the JSON+LD markup from the individual resources needs to be encapsulated in an XML CDATA element. The FIZ OAI-Provider is an open-source software that has been developed for efficiently serving high-volumes of metadata from large resources via the OAI-PMH protocol. The architecture consists of a frontend, backend, the Cassandra document database, and Elasticsearch as an index. For demonstration purposes, we have imported metadata items from the NFDI4Chem resources into a test instance of the OAI-PMH server.

The NFDI4Chem search service is based on CKAN, an open-source data management system, and can harvest metadata through several mechanisms, including an OAI-PMH harvester, as well as JSON+LD harvesting from HTML pages. It has search capabilities relevant in the chemistry domain, including a search for measurement types and molecular structures.

3. Conclusion

We have integrated the capability to serve Bioschemas-compliant metadata into the development versions or local prototypes of several resources in the NFDI4Chem realm, demonstrating that it is possible to load the metadata as JSON+LD into an OAI-PMH server and subsequently harvest into the NFDI4Chem search service through a modified CKAN OAI-PMH harvesting module.

In the future, these developments need to be integrated into the production services in NFDI4Chem. We also invite users with similar requirements to contact us and collaborate to add the approach to the (de facto) standards in metadata management.

One of the main efforts here is the unification of the disparate metadata realms across these resources in chemistry. Obtaining the agreement among all participating resources will take several rounds of prototyping, refining, and implementations. It is not always possible to map the entire data model of the resources to existing schema types, and the agreement might be that, in the case of metadata, some compromises need to be made.

Competing interests

The authors declare that they have no competing interests.

Funding

The consortia are funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the NFDI Funding Programme. The NFDI4Chem project under grant no. 441958208, the NFDI4DataScience project under no. 460234259.

Acknowledgement

The authors acknowledge the BioHackathon Germany 2022 in Lutherstadt Wittenberg, which hosted the "(Bio)Schemas4NFDI" project where some of the ideas were discussed and prototyped.

References

- [1] "Introducing schema.org: Search engines come together for a richer web," *Official Google Blog*. <https://googleblog.blogspot.com/2011/06/introducing-schemaorg-search-engines.html> (accessed Jan. 29, 2023).
- [2] F. Michel and The Bioschemas Community, "Bioschemas & Schema.org: a Lightweight Semantic Layer for Life Sciences Websites," *Biodiversity Information Science and Standards*, vol. 2. p. e25836, 2018. doi: 10.3897/biss.2.25836.
- [3] C. Lagoze and H. Van de Sompel, "The making of the open archives initiative protocol for metadata harvesting," *Libr. Hi Tech*, vol. 21, no. 2, pp. 118–128, Jun. 2003, doi: 10.1108/07378830310479776.
- [4] C. Steinbeck *et al.*, "NFDI4Chem - Towards a National Research Data Infrastructure for Chemistry in Germany," *RIO journal*, vol. 6, p. e55852, Jun. 2020, doi: 10.3897/rio.6.e55852.
- [5] H. Horai *et al.*, "MassBank: a public repository for sharing mass spectral data for life sciences," *J. Mass Spectrom.*, vol. 45, no. 7, pp. 703–714, Jul. 2010, doi: 10.1002/jms.1777.
- [6] S.-A. Sansone *et al.*, "Toward interoperable bioscience data," *Nat. Genet.*, vol. 44, no. 2, pp. 121–126, Jan. 2012, doi: 10.1038/ng.1054.

Terminologies in RDM for Engineering – a Service Approach

NFDI4Ing Terminology Service

Angelina Kraft¹[\[https://orcid.org/0000-0002-6454-335X\]](https://orcid.org/0000-0002-6454-335X), Felix Engel¹[\[https://orcid.org/0000-0002-3060-7052\]](https://orcid.org/0000-0002-3060-7052),
and Axel Klinger¹[\[https://orcid.org/0000-0001-6442-3510\]](https://orcid.org/0000-0001-6442-3510)

¹ Technische Informationsbibliothek (TIB), Germany

Keywords: Semantics, Terminology Service, Ontology, Information Technology

1. Background

The European Commission Expert Group on FAIR Data stated in their action plan in 2018 that “*Semantic technologies are essential for the interoperability and need to be developed, expanded and applied both within and across disciplines*” [1]. As a result, semantic artefacts such as terminologies, ontologies and their respective registries developed among various scientific disciplines. Examples include FAIRsharing [2], BioPortal [3], BARTOC [4], Research Vocabulary Australia [5] and NERC Vocabulary Service [6]. Despite these examples, many disciplines, including large parts of the engineering domain, follow a practice of using ambiguous words, phrases or even incomprehensible abbreviations to annotate data.

To support the coordinated development of RDM services, the National Research Data Infrastructure for Engineering Sciences (NFDI4Ing) provides a Terminology Service. We define a ‘Terminology Service’ (TS) as a web-based platform, which can support the take-up and standardisation of terminologies and ontologies. A controlled terminology thereby is a normative collection of terms whose spelling is fixed and for which additional information such as a definition, synonyms, an editor, a version, and a license can be provided. An ontology on the other hand is a formal representation of the knowledge of a domain, in which concepts are structured and terms are related to each other. A TS may be used in the research data life cycle:

- For findability: Using standardized terminologies, researchers may improve the discoverability of their data;
- For standardization: Terminologies enable researchers to use a common set of terms to describe their data, making it easier to compare and analyse results;
- For integration: Terminologies help to integrate data from different sources by providing a common language for describing data elements and concepts;
- For analysis: Terminologies foster meaningful analyses by ensuring that data is described consistently and unambiguously;
- For sharing and reuse: By using standardized terminologies, researchers can make their data more easily shared and reused by others.

2. The NFDI4Ing Terminology Service

To address the challenge of alignment and reuse of established ontologies and terminologies, a TS was set-up for the NFDI4Ing initiative: <https://terminology.nfdi4ing.de>. The NFDI4Ing TS (**Figure 1**) is a curated resource of terminologies and ontologies for the engineering domain and provides a single point of access to research concepts. In this context, terminologies offer the building blocks for (meta-) data schemata and data annotation. The NFDI4Ing TS enables researchers to browse engineering-related terminologies either through the website or via the Rest API.

The NFDI4Ing TS features more than 50 ontologies, 147,000 terms and over 5,800 properties. An example is the Metadata4Ing (m4i) ontology [7], which enables a process-based description of research activities and their results, focusing on the provenance of both research data and material objects: <https://terminology.tib.eu/ts/ontologies/m4i>. As an open source platform, the NFDI4Ing TS supports the adoption and standardization of ontologies by providing data and knowledge management capabilities for accessing, maintaining, and subscribing to engineering-related terminologies.

The screenshot displays the NFDI4Ing Terminology Service interface. At the top, there is a navigation menu with links for HOME, ONTOLOGIES, HELP, DOCUMENTATION, USAGE, and ABOUT. A search bar contains the text "electric vehicle" and a "Search" button. Below the search bar, there are filter options for "Type" (class, individual, property, ontology) and "Ontologies" (TEMA, OM, BATTINFO). The main content area shows "1617 results found for 'electric vehicle'". Two results are visible: one for "[class] Electric Vehicle" from the schema.mobivoc.org ontology, and another for "[class] electric vehicle" from the openenergy-platform.org ontology. Each result includes a definition and the ontology it belongs to.

Figure 1. The NFDI4Ing Terminology Service provides a single point of access to relevant engineering-related terminologies.

The NFDI4Ing TS provides concepts, terms, relations, their definition and other types of information, from a range of engineering-related terminology collections. Each concept is represented by a Uniform Resource Identifier (URI), which allows the persistent reference to concepts and terms. The NFDI4Ing TS functionalities include a free text search (for- and within ontologies), browsing and filtering, as well as machine-to-machine communication (REST interface). The NFDI4Ing TS is developed and maintained as part of the TIB Central Terminology Service. The Central TS provides other domain-specific terminology collections and is based on the Ontology Lookup Service (OLS) provided by EMBL-EBI [8].

3. Challenges and Outlook

As introduced above, terminologies play a crucial role in ensuring consistency, accuracy, and interoperability of research data. Depending on the research discipline, the availability and quality of terminologies is still limited. With the NFDI4Ing TS, we take a first step to index

available terminologies and ontologies in the engineering domain. The uptake and use of a TS by scientific communities, however, is influenced by many factors and interests, which are more often than not out of the area of influence of the TS providers. The frameworks which have been established since the start of the NFDI provide a chance to align Terminology Services and their management, especially when it comes to questions of quality insurance, versioning, long-term availability, naming conventions, usage of labels and others. Within NFDI, we hope for discussions on these and other topics, and aim to provide trust measures for the NFDI4Ing TS in the future (e.g., labelling of terminology popularity, FAIRness, quality, and new ways of collaborative terminology creation).

Data Availability Statement

The NFDI4Ing Terminology Service including all referenced terminologies is available at <https://terminology.nfdi4ing.de>.

Author Contributions

Angelina Kraft: Conceptualization; Writing – initial draft. Felix Engel and Axel Klinger: Methodology; Supervision; Writing – review and editing.

Competing Interests

The authors declare that they have no competing interests.

Funding

NFDI4Ing Terminology Service is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grant no. 442146713.

Acknowledgement

We thank all PI's and developers which contribute to the NFDI4Ing TS and its main service, the TIB Central Terminology Service.

References

1. European Commission, "Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data", 2018, doi: <https://data.europa.eu/doi/10.2777/1524>
2. S. A. Sansone, P. McQuilton, P. Rocca-Serra, et al. "FAIRsharing as a community approach to standards, repositories and policies". *Nat Biotechnol* 37, 358–367, 2019, doi: <https://doi.org/10.1038/s41587-019-0080-8>
3. P.L. Whetzel, N.F. Noy, N.H. Shah, et al. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications". *Nucleic Acids Res*, 39:W541-5, 2011, doi: <https://doi.org/10.1093/nar/gkr469>
4. Basic Register of Thesauri, Ontologies & Classifications (BARTOC). <https://bartoc.org/> (date accessed: 17 April 2023)
5. Australian Research Data Commons (ARDC). Research Vocabularies Australia. <https://vocabs.ardc.edu.au/> (date accessed: 17 April 2023)
6. British Oceanographic Data Centre. The NERC Vocabulary Server. Natural Environment Research Council. <https://vocab.nerc.ac.uk> (date accessed: 17 April 2023)

7. S. Arndt, B. Farnbacher, M. Fuhrmans, et al., "Metadata4Ing: An ontology for describing the generation of research data within a scientific activity". (1.1.0). Zenodo. doi: <https://doi.org/10.5281/zenodo.7706017>
8. S. Jupp, T. Burdett, C. Leroy, H.E. Parkinson, "A new Ontology Lookup Service at EMBL-EBI". In: Malone, J. et al. (eds.) Proceedings of SWAT4LS International Conference, 2015.

RDM Services at Luxembourg National Data Service

Wei Gu¹[\[https://orcid.org/0000-0003-3951-6680\]](https://orcid.org/0000-0003-3951-6680), Christophe Trefois¹[\[https://orcid.org/0000-0002-8991-6810\]](https://orcid.org/0000-0002-8991-6810), Pinar Alper¹[\[https://orcid.org/0000-0002-2224-0780\]](https://orcid.org/0000-0002-2224-0780), Danielle Welter¹[\[https://orcid.org/0000-0003-1058-2668\]](https://orcid.org/0000-0003-1058-2668), Yohan Jarosz¹[\[https://orcid.org/0000-0003-2047-0897\]](https://orcid.org/0000-0003-2047-0897), Jacek Lebioda¹[\[https://orcid.org/0000-0002-9449-7999\]](https://orcid.org/0000-0002-9449-7999), Linda Ebermann¹[\[https://orcid.org/0000-0002-0862-5561\]](https://orcid.org/0000-0002-0862-5561), Regina Becker¹[\[https://orcid.org/0000-0002-6711-8375\]](https://orcid.org/0000-0002-6711-8375), Venkata Satagopam¹[\[https://orcid.org/0000-0002-6532-5880\]](https://orcid.org/0000-0002-6532-5880), Reinhard Schneider¹[\[https://orcid.org/0000-0002-8278-1618\]](https://orcid.org/0000-0002-8278-1618) and Bert Verdonck¹

¹ Luxembourg National Data Service (PNED G.I.E.) and ELIXIR Luxembourg, Luxembourg

Abstract. The Luxembourg National Data Service (LNDS) was established on 28 July 2022 by the Luxembourg government and public research institutes. LNDS is a national organisation providing services for value creation from public sector data from different domains. LNDS' main mission is to offer technology and data services, know-how, capabilities, platform, and infrastructure to enable sharing and re-use of data for public and private data partners. LNDS aims to become a key component of Luxembourg's data innovation strategy by (1) supporting research and innovation through a high-quality service offering, (2) contributing to accelerating the development of Luxembourg towards a data economy, (3) enabling the participation of both public research and private actors in the complete value-creating chain of data, (4) enhancing data in the context of common innovation among all actors, and (5) becoming a partner of choice for joint development of the next generation of data products.

Keywords: Data stewardship, Community building, Data service

1. Background of LNDS

LNDS is a brand of the Plateforme Nationale d'Échange de Données (PNED G.I.E), an economic interest group. It is positioned to be the competent body of Luxembourg as defined in the EU Data Governance Act (DGA)¹ to support the secondary use of data from the public sectors.

Part of the LNDS' mission is developing a national research data management (RDM) community and supporting the development of RDM capacity in partner institutes through community-driven initiatives. LNDS will mobilise the national RDM community for the development of national RDM guidelines. Another key goal for the LNDS is the elicitation of professional roles for RDM practitioners. We will provide the first formal definitions for the roles of data stewardship and management to inform national stakeholders in building their RDM support workforce.

2. Service development at LNDS

LNDS is designing and building a service portfolio covering the different aspect of data reuse:

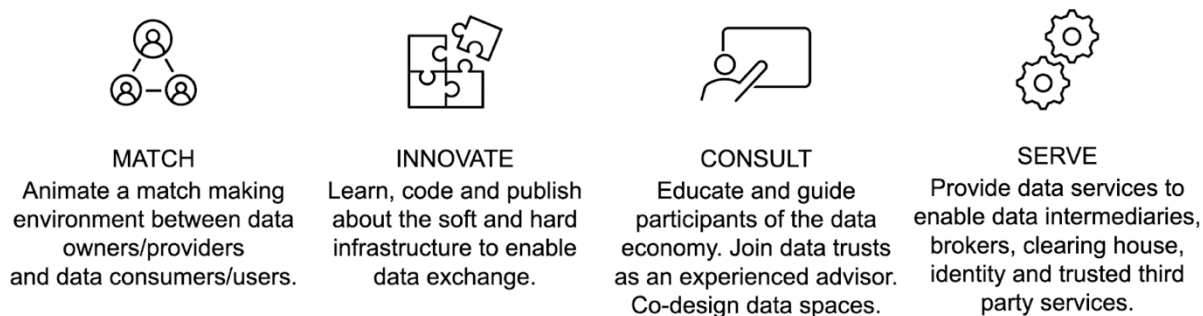


Figure 1. Type of services being developed at the Luxembourg National Data Service.

Through its service, the LNDS also aims to support research data management at research institutes in Luxembourg. The RDM support will cover the life cycle of the research data with a focus on data stewardship, FAIR data, and data sharing.

LNDS' RDM services include Data Management Planning (DMP) support through national tool deployments and templates as well as data management, data stewardship and data protection training. Services for the responsible sharing and re-use of sensitive data such as personal data subject to the GDPR. As part of GDPR support, the services include Data Protection Impact Assessment (DPIA) for prospective data sharing and re-use activities. Implementing controlled access processes for sensitive data, and finally a Trusted Third Party (TTP) service for the primary and secondary pseudonymisation of health and research data.

In addition to the "classical" RDM services, LNDS is also developing services related to other aspects of secondary use of (public sector) data for research and innovation, within the scope of the DGA¹ and the European Health Data Space (EHDS)². This includes data brokerage, secure processing environment, synthetic data, managing data access requests, etc.

LNDS is part of the ELIXIR Luxembourg Node and is represented on the editorial boards of ELIXIR's RDMkit and FAIRCookBook resources. We are also a consortium partner in EU funded projects: European Genomics Data Infrastructure (GDI), leading the Pillar 1/WP2 (long-term sustainability) and WP4 (European operation); Beyond 1 Million Genomes (B1MG), leading WP2 (Ethical, Legal and Social Issues); and the HealthData@EU Pilot (EHDS2-pilot) project.

At the CoRDI 2023 conference, we will introduce the LNDS data management and stewardship services with examples from ongoing partner projects. We will present our first year retrospective of building a national data service for Luxembourg and share experiences and lessons learned.

Data availability statement

The submission is not based on data.

Underlying and related material

Not applicable.

Author contributions

All authors contributed to the conceptualization of the work. WG, CT, PA contributed to the writing – original draft, all authors contributed to the writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Acknowledgement

Not applicable.

References

1. "Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act)". ELI: <http://data.europa.eu/eli/reg/2022/868/oj>
2. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0197>

Swiss-AL: Platform for Language Data in Applied Sciences

On challenges in the field of Language Open Research Data

Julia Krasselt¹[\[https://orcid.org/0000-0003-1060-2657\]](https://orcid.org/0000-0003-1060-2657), Philipp Dreesen¹[\[https://orcid.org/0000-0001-5291-2798\]](https://orcid.org/0000-0001-5291-2798), Peter Stücheli-Herlach¹[\[https://orcid.org/0000-0002-3560-7182\]](https://orcid.org/0000-0002-3560-7182), Dolores Lemmenmeier¹[\[https://orcid.org/0000-0003-0541-6956\]](https://orcid.org/0000-0003-0541-6956), Sooyeon Cho¹[\[https://orcid.org/0009-0005-4172-7008\]](https://orcid.org/0009-0005-4172-7008), Klaus Rothenhäusler¹[\[https://orcid.org/0000-0003-4744-3362\]](https://orcid.org/0000-0003-4744-3362), and Matthias Fluor¹[\[https://orcid.org/0000-0002-0780-8024\]](https://orcid.org/0000-0002-0780-8024)

¹ ZHAW Zurich University of Applied Sciences

Abstract. Open Science is transforming the way researchers collect, process, analyze, and store empirical research data, particularly in the social sciences and humanities, where language data is crucial. This transformation process especially concerns developers and providers of large language corpora and manifests itself in at least three challenges when providing these corpora as Open Research Data (ORD). Challenges concern heterogeneous practices that researchers apply when working with language data, research data lifecycle, and legal and ethical aspect. In this paper, we present Swiss-AL, a language data platform developed in Switzerland that is being transformed into an Open Research Data Resource for Applied Sciences within the Swiss Open Science Strategy. The paper gives an overview over the data contained in Swiss-AL and the infrastructure that is used to process and analyze the data. Furthermore, it presents approaches to the three abovementioned challenges to language ORD.

Keywords: Language Data, Corpus Linguistics, Interdisciplinarity

1. Introduction

Open Science is revolutionizing the way researchers collect, process, analyze, and store empirical research data, particularly in the social sciences and humanities, where language data is crucial. However, sharing large language corpora with a diverse research community presents unique challenges, including structured and FAIR data access, disciplinary research practices, and compliance with copyright and data protection laws. Here, We will introduce Swiss-AL, a language data platform for Applied Sciences developed at ZHAW, Switzerland, that is currently being developed into an Open Research Data (ORD) Resource for the Swiss and European CLARIN Community This paper presents Swiss-AL's approach to these challenges, which is currently funded under the Swiss Open Science Strategy.

2. Swiss-AL: Platform for Language Data

2.1 Corpora and Processing Pipeline

Swiss-AL is a multilingual text collection designed for analyzing public communication and relevant societal discourses in interdisciplinary and transdisciplinary contexts [1]. The corpus currently contains 4.5 billion words, making it the largest Swiss text collection [2].

Swiss-AL provides three types of text collections: Swiss-AL Base, Swiss-AL Media, and Swiss-AL Projects (Fig. 1). Swiss-AL Base contains web-crawled texts published by various public actors, including political and administrative authorities, industry associations, Swiss universities, civil society, and newspapers. Swiss-AL Media focuses exclusively on journalistic media from major and regional newspapers of Swiss publishing houses. Swiss-AL Projects consists of thematically specific corpora compiled for research projects and shared with the research community.

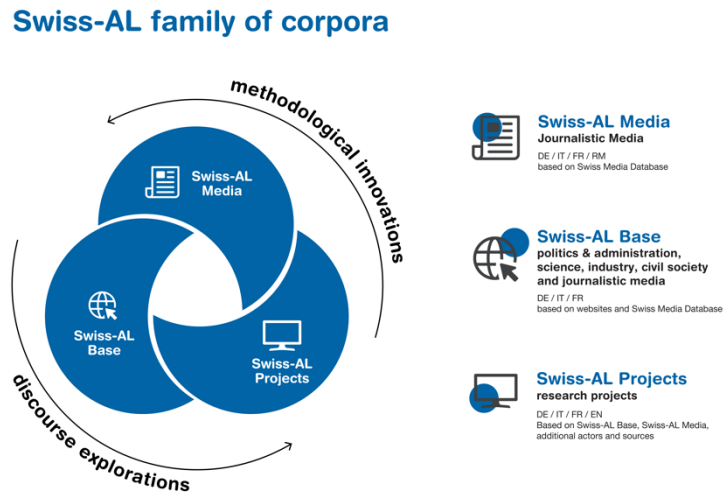


Figure 1. Swiss-AL family of corpora

Swiss-AL corpora are compiled using a computational linguistic pipeline that can adapt to different types of input data. The focus is on dynamic parts of websites (subpages covering news reports, media releases, blogs), using a web crawler and web page-specific Xpath scrapers. The data is loaded into an Elastic Search database in a structured form. Filters are applied to process texts in language-specific ways and to recognize and sort out near-duplicates [3]. The linguistic processing is based on the UIMA framework [4] and various modules such as TreeTagger for PoS tagging [5], Stanford NER for named entity recognition [6], the Stanford Dependency Parser [7] or in-house developed tools for additional annotation layers.

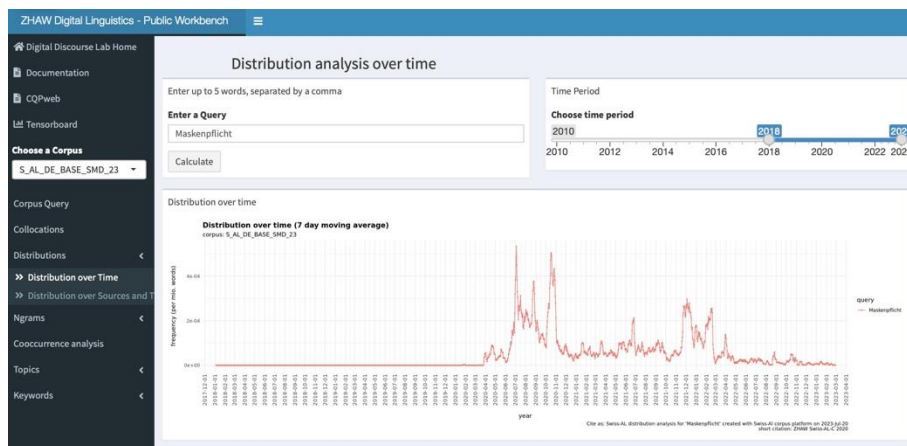


Figure 2. Swiss-AL Workbench. Left: Selection of a sub-corpus and a desired analysis method, access to the documentation. Right: Display of the results. Here, the frequency of a word (*Maskenpflicht*) over time (2018-2023) in the Swiss-AL-Base corpus is shown.

2.2 Access: Swiss-AL Workbench

Swiss-AL corpora can be accessed via an in-house developed public web app, the *Swiss-AL Workbench* (Fig. 2, [8]). The focus is on aggregating methods of data analysis (*distant reading*), in which phenomena on the language surface are summarized quantitatively. The workbench enables word-based query methods (e.g., word distributions over time) and analyses using machine and deep learning methods (topic modelling, word embeddings). In the course of developing Swiss-AL into an ORD resource, the workbench will be reimplemented by the end of 2024 in order to approach the challenges mentioned in the following section.

3. Challenges related to Linguistic Open Research Data

Text data not only form the empirical basis in linguistics, but in many other research disciplines. This makes language data like Swiss-AL different from other data like genome sequences and particularly valuable as an ORD resource. At the same time, however, this also gives rise to special challenges, three of which we will discuss in the following sections.

3.1 Heterogeneous research practices

Humanities, social sciences, communications sciences, law, and architecture studies are examples for disciplines interested in language data, each employing diverse research practices with labels such as *qualitative*, *quantitative*, *data-driven*, *hypothesis-driven*, *close reading*, and *distant reading*. When developing Swiss-AL into an ORD resource, it is crucial to consider these varied user groups and research methods. These users, labelled semi-professionals from a corpus-linguistic perspective, are experienced empirical researchers in their fields but unfamiliar with corpus linguistic methods and linguistic surface analysis. Keeping this target user group in mind, we developed user stories, which answer who wants to do what for which purpose. Acceptance criteria derived from these user stories guide the re-implementation of the existing workbench.

3.2 Research Data Lifecycle

Billion-word corpora like Swiss-AL need special infrastructures for structured access to primary data, metadata, and documentation, as conventional repositories risk violating copyright law and serve only a limited expert community. Swiss-AL promotes reuse by integrating curated data, storage, and analysis tools to support interdisciplinary communities via a dynamic FAIR infrastructure. Traditional repositories, found at the data lifecycle's end, are less conducive to reuse due to the divide between storage and analysis tools. FAIR principles are essential for re-implementing the Swiss-AL workbench. E.g., it will enable users to perform not only word-based analysis, but also to download corresponding data frames, corpus information, and code for reproducing visualizations.

3.3 Legal and ethical aspects

Publishing language data as an ORD resource requires considering legal and ethical aspects. Texts from SMD and crawled web data in Swiss-AL are protected by Swiss copyright law. However, under §24d of the Swiss Copyright Act, reproducing work for scientific research is permissible if it involves a technical process and the copied work is lawfully accessible. Thus, crawling web data for research is allowed. However, representing full texts, a common researcher need, warrants caution. A scientific legal opinion is currently being prepared in collaboration with lawyers to examine possibilities in this domain.

Secondly, language corpora such as Swiss-AL contain personal data (e.g., journalistic media articles mentioning real persons), i.e., the identification of an individual person is potentially possible. Thus, data protection law needs to be considered when obtaining, saving, and

publishing corpus data. The topic is well known from other empirical research fields, e.g., when conducting qualitative interviews or doing field work. However, the practice of anonymisation typically used there is not a practicable solution for large linguistic corpora. In particular, an envisaged solution is to formulate a purpose for which the data contained in Swiss-AL will be collected, stored and analysed.

4. Conclusion

Language data are not prototypical ORD, and they are not an exclusive data resource for linguistics. However, if language data are to be made available as ORD for other disciplines, a variety of challenges arise. These can only be solved in an interdisciplinary way, taking into account technical, legal and ethical aspects on a societal and international level.

Data availability statement

The corpora described in the article can be accessed under www.swiss-al.linguistik.zhaw. A documentation of the corpora is available under <https://swiss-al.linguistik.zhaw.ch/docs/ord/>.

Author contributions

Conceptualization: JK, PD, PSH; Writing – original draft: JK; Writing – review & editing: JK, PD, PSH, DL, SC; Software: KR, MF

Competing interests

The authors declare that they have no competing interests.

Funding

The project presented in this paper is currently funded by swissuniversities within the programme Swiss Open Research Data Grants (CHORD) and the Zurich University of Applied Sciences (internal funding).

References

- [1] P. Dreesen and P. Stücheli-Herlach, "Diskurslinguistik in Anwendung. Ein transdisziplinäres Forschungsdesign für korpuszentrierte Analysen zu öffentlicher Kommunikation", *Zeitschrift für Diskursforschung*, vol. 7, no. 2, pp. 123–162, 2019, doi: <https://doi.org/10.3262/ZFD1902123>
- [2] J. Krasselt, P. Dreesen, M. Fluor, C. Mahlow, K. Rothenhäusler, and M. Runte, "Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics", in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, 2020, pp. 4138–4144. <https://aclanthology.org/2020.lrec-1.510/> [26.04.2023]
- [3] M. Theobald, J. Siddharth, and A. Paepcke, "SpotSigs: Robust and Efficient Near Duplicate Detection in Large Web Collections", in *31st annual international ACM SIGIR conference on Research and development in information retrieval 2008 (SIGIR 2008)*, Singapore, Singapore, 2008.

- [4] D. Ferrucci and A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research environment", *Natural Language Engineering*, vol. 10, no. 3–4, pp. 327–348, 2004, doi: <https://doi.org/10.1017/S1351324904003523>.
- [5] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees", in *Proceedings of the international conference on new methods in language processing*, Manchester, United Kingdom, 1994, pp. 44–49. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.1139>
- [6] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, Michigan, 2005, pp. 363–370.
- [7] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit", in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland USA, 2014, pp. 55–60. [Online]. Available: <http://www.aclweb.org/anthology/P/P14/P14-5010>
- [8] J. Krasselt, M. Fluor, K. Rothenhäusler, and P. Dreesen, "A workbench for corpus linguistic discourse analysis", in *3rd conference on language, data and knowledge (LDK 2021)*, D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, and B. Heinisch, Eds., in *Open access series in informatics (OASlcs)*, vol. 93. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, p. 26:1-26:9. doi: <https://doi.org/10.4230/OASlcs.LDK.2021.26>.

MaRDMO Plugin

Document and Retrieve Interdisciplinary Workflows Using the MaRDI Portal.

Marco Reidelbach¹[\[https://orcid.org/0000-0002-1919-1834\]](https://orcid.org/0000-0002-1919-1834), Eloi Ferrer¹[\[https://orcid.org/0009-0009-2327-2619\]](https://orcid.org/0009-0009-2327-2619),
and Marcus Weber¹[\[https://orcid.org/0000-0003-3939410X\]](https://orcid.org/0000-0003-3939410X)

¹Mathematics of Complex Systems, Zuse Institute Berlin, Berlin, Germany

Abstract: MaRDMO, a plugin for the Research Data Management Organiser, was developed in the Mathematical Research Data Initiative to document interdisciplinary workflows using a standardised scheme. Interdisciplinary workflows recorded this way are published directly on the MaRDI portal. In addition, central information is integrated into the MaRDI knowledge graph. Next to the documentation, MaRDMO offers the possibility to retrieve existing interdisciplinary workflows from the MaRDI Knowledge Graph to allow the reproduction of the initial work and to provide scientists with new research impulses. Thus, MaRDMO creates a community-driven knowledge loop that could help to overcome the replication crisis.

Keywords: Interdisciplinary Workflows, RDMO, MaRDI Portal

1 Introduction

Reproducibility is a key element of science. Increasing reports about reproduction problems [1], [2] impair the scientific credibility and limit synergies arising from reusing and advancing scientific work. The reasons for reproducibility issues are manifold, including intentional misbehaviour [3], but also incomplete documentation of applied methodologies and software, missing parameters or conditions, and inconsistent naming schemes [2].

To overcome these issues the Mathematical Research Data Initiative (MaRDI [4]) developed a standardised documentation scheme to summarise the relevant aspects of interdisciplinary workflows independent of the research area and its design (theoretical/experimental) [5]. Thereby, scientists from various disciplines are able to document their interdisciplinary workflows and publish them on the MaRDI portal, a unique portal that will contain all sorts of mathematical research data. Likewise, scientists are able to retrieve documentations through the corresponding Knowledge Graph (KG) to reproduce, develop, and re-document the original work to retain the knowledge in a reproducible form.

To enable scientists from the entire research community to document and retrieve interdisciplinary workflow documentations using the MaRDI portal, a low-threshold access is necessary. This means that the access should ideally be integrated into the

normal working day, without the need to call up the MaRDI portal directly and edit/query it manually.

The Research Data Management Organiser (RDMO) is an open source web application for research data management. It provides discipline-specific question catalogues and allows the output of data management plans. Currently, RDMO is used in many research institutions in Germany and tested in other European countries [6]. It is very likely that use will continue to increase in the coming years, especially since most of the National Research Data Infrastructure consortia have endorsed the use of RDMO [7]. Thus, it is an integral part of the work of many scientists from the entire research community.

Since RDMO encourages the design of individual question catalogues and additional functionalities, a question catalogue for interdisciplinary workflow documentation and a functionality connecting RDMO with the MaRDI portal could establish the desired low-threshold portal access.

Here, we present MaRDMO, a plugin for RDMO, to document and retrieve interdisciplinary workflows using the MaRDI Portal.

2 MaRDMO Plugin

The MaRDMO Plugin contains a questionnaire and an export/query functionality. It adds a "MaRDI Export/Query" button to the project view by which the entire plugin is controlled. The remaining settings, cf. 1, are made via the question catalogue.

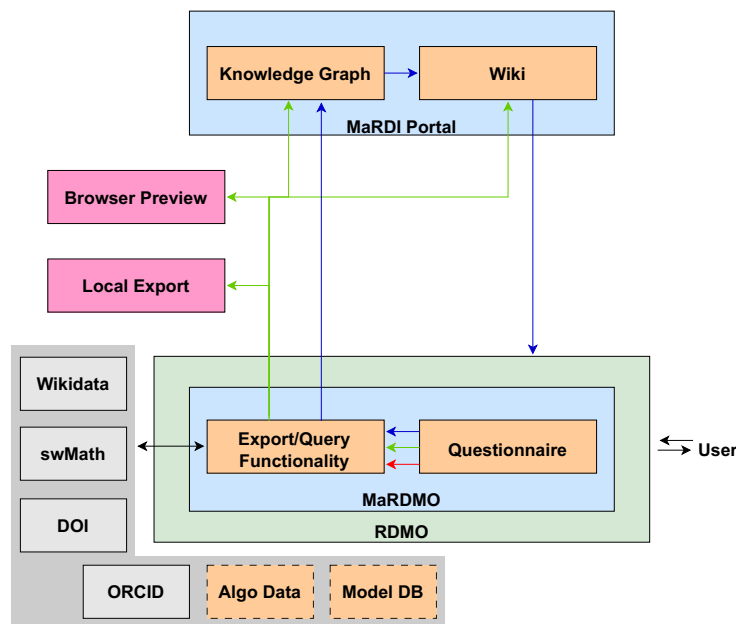


Figure 1. Overview of MaRDMO - MaRDI portal interaction. Interdisciplinary workflow documentation (green) and retrieval (blue) pathways, plugin settings (red) and interactions with other data sources (black) are indicated by arrows. AlgoData and ModelDB are MaRDI services which will be integrated in the future.

2.1 Question Catalogue

The MaRDMO questionnaire is divided in 7 sections, serving three purposes, i.e. documentation or retrieval of interdisciplinary workflows <https://de.overleaf.com/project/64e61490cb3d6285a> and plugin settings.

To do a documentation questions providing a general description, e.g. research objective, involved disciplines and data streams, describing the underlying mathematical model, e.g. model description, involved variables and parameters and discretisation, providing the relevant process information, e.g. process steps, applied algorithms, used software, hardware, experimental devices, input and output data, and describing the reproducibility need to be answered. All entries should be provided with identifiers, e.g. MaRDI identifiers, Wikidata QIDs, DOIs and swmath IDs, to properly integrate the individual documentations into the existing research data landscape.

To retrieve documentations the entities to search for, e.g. research objective, applied model, methods and software, need to be defined. Entities may be searched via keywords or identifiers.

2.2 Export/Query Functionality

The export/query functionality gathers all the information provided via the questionnaire and creates a documentation or retrieves documentations by an entity search.

For documentations, all answers are integrated into MaRDI's predefined documentation scheme. Depending on the user's needs, the completed scheme is then saved locally, previewed in the browser, e.g. to detect faulty rendering, or sent to the MaRDI portal to create a wiki page.

If the documentation is sent to the MaRDI portal, individual components, e.g. name, associated publication, research objective, model, methods, software and input data, are also integrated into the KG of MaRDI.

To exemplarily link a new interdisciplinary workflow item with an associated publication, the following steps are performed:

- Check if an item with the custom DOI already exists in MaRDI KG, and if so, link it to the new interdisciplinary workflow item.
- Check if an item with the custom DOI already exists in Wikidata KG, if so, integrate it into MaRDI KG and link it to the new interdisciplinary workflow item.
- If an item with the custom DOI cannot be found in either KG, get the full citation from DOI and ORCID, create a new item in MaRDI KG, and link it to the new interdisciplinary workflow item. Repeat the first two steps for the authors and journal. If no entries are found, create them and link them to the new publication entry.

To retrieve documentations, all answers are translated into a SPARQL query. If relevant information is found in the MaRDI KG, appropriate documentations are returned as wiki pages.

3 Conclusion

MaRDMO is a software package that enables the entire research community to document and retrieve interdisciplinary workflows using established infrastructure (RDMO). Documentations are integrated into the existing research data landscape to exploit net-

work effects and provide scientists with instant visibility, while the retrieval of documentations brings together new research impulses from various disciplines. Due to its simple structure, MaRDMO may also be used with other questionnaires (and KGs) to document and search research objects other than interdisciplinary workflows.

Data availability statement

The MaRDMO Plugin is available and documented on <https://github.com/MarcoReidelbach/MaRDMO>.

Underlying and related material

Knowledge Graph settings (items and properties) required for MaRDMO and a script to make any Wikibase Knowledge Graph compatible with MaRDMO can be found on <https://github.com/MarcoReidelbach/MaRDMO>.

Author contributions

Marco Reidelbach (Conceptualization, Software, Methodology, Writing - Original Draft, Writing - review & editing), Eloi Ferrer (Software, Methodology, Writing - review & editing) and Marcus Weber (Conceptualization, Writing - review & editing, Funding acquisition)

Competing interests

The authors declare that they have no competing interests.

Funding

Marco Reidelbach, Eloi Ferrer and Marcus Weber are supported by MaRDI, funded by the Deutsche Forschungsgemeinschaft (DFG), project number 460135501, NFDI 29/1 "MaRDI – Mathematische Forschungsdateninitiative".

Acknowledgements

We thank MaRDI's Task Areas 4 and 5 for fruitful discussions concerning workflow documentation and portal integration.

References

- [1] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, pp. 452–454, May 2016. DOI: [10.1038/533452a](https://doi.org/10.1038/533452a).
- [2] Krishna Tiwari, Sarubini Kananathan, Matthew G. Roberts, Johannes P. Meyer, Mohammad Umer Sharif Shohan, Ashley Xavier, Matthieu Maire, Ahmad Zyoud, Jinghao Men, Szeyi Ng, Tung V. N. Nguyen, Mihai Glont, Henning Hermjakob and Rahuman S. Malik-Sheriff, "Reproducibility in systems biology modelling," *Molecular Systems Biology*, vol. 17, e9982, Feb. 2021. DOI: [10.15252/msb.20209982](https://doi.org/10.15252/msb.20209982).
- [3] D. Fanelli, "How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data," *PLOS ONE*, vol. 4, no. 5, e5738, May 2009. DOI: [10.1371/journal.pone.0005738](https://doi.org/10.1371/journal.pone.0005738).

- [4] Christiane Görge and Rainer Sinn, "Mathematik in der nationalen forschungsdateninfrastruktur," *Mitteilungen der Deutschen Mathematiker-Vereinigung*, vol. 29, no. 3, pp. 122–123, Oct. 2021. DOI: [10.1515/dmvm-2021-0049](https://doi.org/10.1515/dmvm-2021-0049).
- [5] Tobias Boege, René Fritze, Christiane Görge, Jeroen Hanselman, Dorothea Iglezakis, Lars Kastner, Thomas Koprucki, Tabea Krause, Christoph Lehrenfeld, Silvia Polla, Marco Reidelbach, Christian Riedel, Jens Saak, Björn Schembera, Karsten Tabelow and Marcus Weber, "Research-data management planning in the german mathematical community," Nov. 2022. DOI: [10.48550/arXiv.2211.12071](https://doi.org/10.48550/arXiv.2211.12071).
- [6] R. Arbeitsgemeinschaft. "RDMO - Cooperations." (2023), [Online]. Available: <https://rdmorganiser.github.io/en/cooperations/> (visited on 04/16/2023).
- [7] Harry Enke, Daniela Hausen, Christin Henzen, Gerald Jagusch, Celia Krause, Sabine Schönau, Lukas Weimer and Jürgen Windeck, "Data management planning: Concept for setting up a working group in the nfdi section common infrastructures," Jan. 2023. DOI: [10.5281/zenodo.7540682](https://doi.org/10.5281/zenodo.7540682).

Metadata Fields and Quality Criteria - XAS Reference Database under DAPHNE4NFDI

Abhijeet Gaur^{1*}, Sebastian Paripsa³, Frank Förste⁴, Dmitry Doronkin^{1,2}, Wolfgang Malzer⁴, Christopher Schlesiger⁴, Birgit Kanngießner⁴, Dirk Lützenkirchen-Hecht³, Edmund Welter⁵, and Jan-Dierk Grunwaldt^{1,2}

¹ Institute for Chemical Technology and Polymer Chemistry, Karlsruhe Institute of Technology (KIT), Engesserstr. 20, Karlsruhe, D-76131 (Germany)

² Institute of Catalysis Research and Technology, Karlsruhe Institute of Technology (KIT), Hermann-von-Helmholtz-Platz 1, Eggenstein-Leopoldshafen, D-76344 (Germany)

³ Fk. 4, Physik, Bergische Universität Wuppertal, Gauß-Str. 20, Wuppertal, D-42097 (Germany)

⁴ Technische Universität Berlin, Hardenbergstr. 36, Berlin, D-10623

⁵ Deutsches Elektronen-Synchrotron (DESY), Notkestraße 85, Hamburg, D-22607

Abstract. X-ray absorption spectroscopy is important to analyse solid materials, in particular amorphous materials, disordered or multicomponent materials. Due to its vast application in diverse scientific fields XAS has become an essential tool for studying, e.g., catalytic reactions or battery materials to mention just a few. In the field of XAS, data are often evaluated by comparing them to previously measured or calculated reference spectra. This sets the high requirements concerning both spectral quality and documentation of the measurements. Under DAPHNE4NFDI, we have been working on to set up a XAS reference database including raw and processed data with an interface developed for uploading and evaluating the data. In this context, defining metadata fields about the performed XAS experiments and documenting this information along with data is essential to make the measured data reusable by any researcher in a similar field and beyond. Another important aspect of a curated database is that users should be able to easily judge the quality and the usability of each data set by looking at the mentioned quality criteria. In the present work, we have discussed and highlighted the importance of metadata fields and quality criteria for the data to be uploaded at the XAS database.

Keywords: X-ray absorption spectroscopy, Database, Metadata, Quality criteria

1. Extended Abstract

X-ray absorption spectroscopy is important to analyse solid materials, in particular amorphous materials, disordered or multicomponent materials. Due to its vast application in diverse fields XAS has become an essential tool for studying, e.g., catalytic reactions or battery materials to mention just a few. In the field of XAS, data are often evaluated by comparing them to previously measured or calculated reference spectra [1]. This sets the high requirements concerning both spectral quality and documentation of the measurements. Previous databases for XAS suffer from problems such as not providing detailed information about the sample and the data acquisition process itself, unknown and inconsistent data formats, difficulties in adding new data to the database, non-standardized organisation of the database, and a review process to ensure and assess the quality of submitted data [2]. Recently, XAFS databases in Japan have

been integrated to create a new database called MDR XAFS database [3], which hosts approximately spectra of 2000 samples and 700 unique materials with machine-readable metadata. However, data quality is not included in the inclusion criteria for this collection and hence use of data is at the discretion of the user.

Under DAPHNE4NFDI [4], we have been working on to set up a XAS reference database including raw and processed data with an interface developed for uploading and evaluating the data. In this context, defining metadata fields about the performed XAS experiment and documenting this information along with data is essential to make the measured data reusable by any researcher in a similar field and beyond. For the present database, we have categorized meta data fields under "Sample", "Spectra", "Instrument" and "Bibliography", sub-fields under these categories are shown in Scheme 1. Hence, the metadata fields include contributions from users as well as experimental facilities [5]. Another important aspect of a curated database is that users should be able to easily judge the quality and the usability of each data set by looking at the mentioned quality criteria. Quality criteria, i.e., edge step, energy resolution, signal to noise ratio, have been considered for the automated check of any uploaded data. The edge step is directly related to the elemental composition of the sample, the concentration of elements, as well as sample preparation. The energy resolution is generally dependent on the limits of the used beamline/instrument. The assigned values of quality criteria shown in Scheme 2 corresponds to references, e.g., metal foils. During the upload of the data, an automatic check of these defined quality criteria and noise estimation is performed. Details about data upload and quality check procedures are available at our database public webpage [6].

Scheme 1. Defined metadata fields for the XAS database

Sample

- Collection code
- CAS no. (optional)
- Physical state (crystalline, powder, thin film, liquid, gas)
- Structural parameters for crystalline sample
- Crystal orientation
- X-ray or neutron diffractogram (if available)
- Temperature
- Pressure
- Remarks about sample preparation – Foil, Pellet, Capillary, Powder on tape, etc.
- Sample environment - Cell/Microreactor/Batch, gases, solvents, potential, etc.
- General remarks e.g., sample properties (hygroscopic metastable, etc).

Bibliography

- DOI
- Title
- Author
- Reference
- Funding agency /Grant details

Spectra

- Raw data file (optional)
- Absorbance spectrum (3000 data points)
- Reference spectrum
- Data/File format
- Header information

Instrument

- Facility (synchrotron/lab instrument)
- Beamline
- Acquisition mode
- Crystals
- Mirrors
- Harmonic rejection
- Detectors
- Element
- Edge
- Maximum k range (EXAFS)

The metadata fields and quality criteria are still an open point of discussion to cover the different types of experiments (ex situ, in situ, operando), acquisition modes, instruments (synchrotron beamline/laboratory facility), detection modes, data formats etc. The database will be filled with real spectra covering a wide range of functional materials, with criteria for meta-data and quality assessment. It will allow researchers to easily access and analyze XAS data. The interface for data submission will make it easy for users to contribute their own data

to the database, and the automated assessment of data quality will ensure that the data in the database is of high quality. This will not only benefit the researchers who contribute data, but also other researchers who use the database for their own work.

Scheme 2. Formulated quality criteria for uploaded data (values corresponds to metal foils)

Quality criteria

- edge step [0.5 - 2.0]
- maximum k-range [15 - 20 Å⁻¹]
- energy resolution [0.5 - 2.0 eV]
- signal to noise ratio [comparison procedure]
- amplitude reduction factor [0.7 - 1.0]*

*analysed data

One of the application of such curated database is that it would be possible to compare the data for identical samples from different facilities and hence the effect of different parameters of an instrument on the data quality can be studied. As an example Fig. 1 shows the comparison of Pt foils measured at XAS beamlines from different synchrotron facilities. However, these Pt foils may not be identical, e.g., thickness, purity etc., hence their data is also affected by these factors other than beamline parameters. Thus, identical samples need to be measured at different XAS facilities (synchrotron/laboratory) by using either their own or standardized analytical protocols. This is basic idea behind a recently initiated Round Robin test [7] which could further help to standardize the meta data fields across different laboratories.

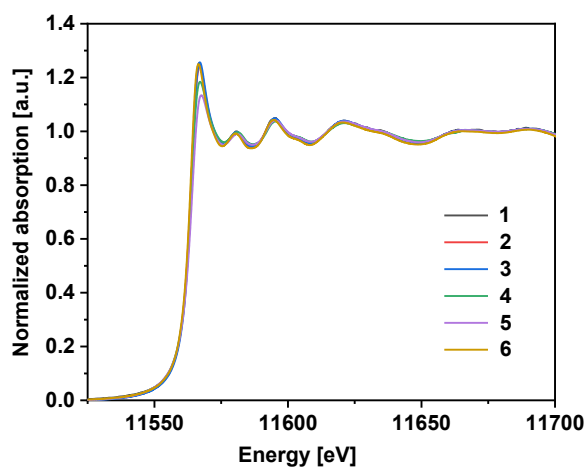


Figure 1. Comparison of Pt foils (distinct) data measured at different beamlines.

As such, our objective is to establish a database for X-ray Absorption Spectroscopy (XAS) data in a timely manner to create a self-accelerating effect and extend the knowledge to other fields. Our overarching goal under DAPHNE4NFDI is to develop and endorse efficient and systematic data and metadata capturing tools during experiments, establish appropriate metadata schemata for respective communities, create federated data catalogues and repositories, and provide tools for data processing, visualisation, and analysis.

Underlying and related material

--

Author contributions

Please include a statement on authors' contributions according to the [CreDIT guidelines](#) here. CRediT (Contributor Roles Taxonomy)'s intention is to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Competing interests

The authors declare that they have no competing interests.

Funding

Deutsche Forschungsgemeinschaft <http://dx.doi.org/10.13039/501100001659> 460248799

Acknowledgement

This work was supported by the consortium DAPHNE4NFDI (<https://www.daphne4nfdi.de/>) in the context of the work of the NFDI e.V.

References

1. A. Gaur and B. D. Shrivastava, "Speciation using X-ray absorption fine structure (XAFS)" *Rev. J. Chem.*, 5, pp. 361-398, November, 2015, doi: <https://doi.org/10.1134/S2079978015040032>
2. K. Asakura, H. Abe., M. Kimura, "The challenge of constructing an international XAFS database" *J. Synchrotron Rad.*, 25, pp. 967-971, July, 2018, doi: <https://doi.org/10.1107/S1600577518006963>
3. M. Ishii, K. Tanabe, A. Matsuda, H. Ofuchi, T. Matsumoto, T. Yaji, Y. Inada, H. Nitani, M. Kimura and K. Asakura, "Integration of X-ray absorption fine structure databases for data-driven materials science" *Sci. Technol. Adv. Mater.: Methods*, 3, pp. 2197518, doi: <https://doi.org/10.1080/27660400.2023.2197518>
4. DAPHNE4NFDI - Consortium Proposal <https://doi.org/10.5281/zenodo.8040606>
5. R. Dimper, A. Götz, A. De Maria, M. Solé V.A., Chaillet, B. Lebayle, "ESRF Data Policy, Storage, and Services", *Synchrotron Radiation News*, 32, pp.7-12, May, 2019, doi: <https://doi.org/10.1080/08940886.2019.1608119>
6. <https://san-wierpa.github.io/xafsdb4daphne/>
7. C. T. Chantler, B. A. Bunker, H. Abe, M. Kimura, M. Newville, E. Welter, "A call for a round robin study of XAFS stability and platform dependence at synchrotron beamlines on well defined samples", *J. Synchrotron Rad.*, 25, pp. 935-943, July, 2018, doi: <https://doi.org/10.1107/S1600577518003752>

Opportunities and Limits of a Disciplinary Repository Using the Example MO|RE data (eResearch Infrastructure for Motor Research Data)

Katja Klemm¹[\[https://orcid.org/0000-0002-8491-2247\]](https://orcid.org/0000-0002-8491-2247), Hannah Kron¹[\[https://orcid.org/0000-0001-5733-2741\]](https://orcid.org/0000-0001-5733-2741), Alexander Woll¹[\[https://orcid.org/0000-0002-5736-2980\]](https://orcid.org/0000-0002-5736-2980), Klaus Bös¹[\[https://orcid.org/0009-0002-9195-7133\]](https://orcid.org/0009-0002-9195-7133) and Claudia Niessner¹[\[https://orcid.org/0000-0003-2094-0836\]](https://orcid.org/0000-0003-2094-0836)

¹ Karlsruhe Institute of Technology, Germany

Abstract. Out of two funding periods by the German Research Foundation resulted the first disciplinary repository for sports science motor activity research data MO|RE data. MO|RE data addresses sports scientists, researchers from related disciplines and practitioners such as teachers and educators, which work with or generate human motor performance test data as well. It has five main functions: publishing, storing, searching, citing and mapping. There are still some limitations as well as not exploited opportunities. Opportunities include among others the international expansion. Limitations are e.g., that linking health and motor test data is difficult due to data protection law. The sharing of sensitive data is not possible with the current concept and needs further solutions as e.g., remote access or workstations for guest researchers. In future, those data sets must be kept in mind to cover the requirements of research data management in sports science. Overall, the need for further development and optimization of MO|RE data repository in sports science becomes apparent to maximize its potential and ensure that it meets the evolving needs of researchers in the field.

Keywords: Sport Science, Open Data, Motor Tests

1. Introduction

The Institute of Sports and Sport Science (IFSS) at the KIT is one of the biggest research centres for human motor performance testing. Meanwhile, the institute collected over 250 000 data points during (partner) projects and published 29 test profiles for different target groups, settings and motor abilities. The mostly used one is the German Motor Test for children and youth from 6 to 18 years. Until 2013, there existed no data management, storage and publication solution in sport science or regarding motor test data in general. To tackle this lack, the IFSS started the project “eResearch infrastructure for motor research data” in 2014. Partner of this project is the service team RDM at the KIT library.

Out of two funding periods by the German Research Foundation resulted the first disciplinary repository for sports science motor activity research data MO|RE data. This work displays its main functions, the upcoming opportunities and possible limitations.

2. Background

Different studies examined a great willingness to share own data and/or a great interest in using “foreign” data in different scientific disciplines [1-3]. Kloe et al. did a demand analysis in sport science in 2019. They found out, that 81.7% of sports scientists in German-speaking

countries are interested in data sharing. Out of sports scientists with own generated data it was even 91.5%. [4]

Motor test data contains numbers related to age and sex (e.g., how many push-ups achieves an 8-year-old boy?). The mostly used related data (called "additional data") are e.g. BMI, physical activity and further health or fitness values. This means the data sets in MO|RE data are very homogenous and, in relation to data sets of nature sciences, very small and need little storage.

3. Functions of MO|RE data

MO|RE data addresses sports scientists, researchers from related disciplines and practitioners such as teachers and educators, which work with or generate human motor performance test data as well. Additionally, practitioners shall be encouraged to use MO|RE data as an information platform, e.g. to get an impression of the current research state or to compare their data (e.g. of a school class) with a scientific based data set.

For reaching those target groups different functions need to be combined. Therefore, MO|RE data provides the following functions:

- Publishing: Raw and aggregated data sets and the according metadata can be published via MO|RE data using a Creative Commons license (CC BY-SA or CC-BY). Therewith the data set is visible for other users and can be downloaded for reuse.
- Mapping: When uploading a data set, users can match variables of their file with prepared variables in MO|RE data.
- Citing: Data sets can be identified with the digital object identifier (DOI). Every data set receives an identifier when getting published.
- Storing: The publication involves the archiving of the data sets, which presents a sustainable and safe option for users to store their own data.
- Searching: In MO|RE data, users can search for test items, age, gender, authors and much more variables in single or combined search terms.

4. Opportunities and limitations

With its aims, functions and background, MO|RE data presents opportunities as well as limitations in the further development.

4.1 Opportunities

At first sight, MO|RE data is a data repository for sports scientists to archive and publish their motor test data. At the second glance, there exist much more opportunities, which are already available or can be developed with small to medium-level effort.

Within the current functions, scientists from other disciplines can upload their data as well. Big interdisciplinary questions as the correlation between motor abilities and school performance or motor abilities and health issues can be answered by bringing together the data sets. Supposedly small data sets can provide new insights when matched with others.

Furthermore, the repository is based on national and international research. Test items were chosen for their usage in Germany and abroad, to build a basis for international cooperations on data sharing.

Challenging is, how to communicate these opportunities to the according target group and to match them with the existing limitations, following in the next paragraph.

Finally, these functions can be expanded with small effort: (1) the mapping variables can be extended, for example if a related discipline wishes further group variables, (2) the search function can be developed and specified.

4.2 Limitations

Due to data protection laws, open data access does not allow other important health data (BMI, blood pressure) or information on personal data (geolocation data, social status) to be published in MO|RE data because that would make it possible to identify an individual person. However, linking the health data with the physical fitness data and follow-up data (longitudinal data sets) are of great scientific interest.

The data and information mentioned above are already available from studies, but only a very small part can be made publicly available in the database MO|RE data. The exchange of data with interested researchers therefore currently takes place mostly in person and often requires long and expensive journeys, or sometimes does not take place at all due to the high effort involved.

A protected digital space for data exchange would make it possible to link the physical fitness test data with other important data, e.g., on health, and to make these available to interested researchers while strictly maintaining data protection.

5. Discussion

With MO|RE data the first repository in sports science exists, but there are still some limitations as well as not exploited opportunities. The sharing of sensitive data is not possible with the current concept and needs further solutions as e.g., remote access or workstations for guest researchers. In future, those data sets must be kept in mind to cover the requirements of research data management in sports science. In addition, network building and establishment in the field and in the related disciplines are essential to promote sustainability. By fostering collaborations and building networks, the repository can continue to expand, providing a valuable resource for researchers in the field of sports science.

Data availability statement

This submission is not based on data.

Author contributions

All named authors contributed to this abstract: KK wrote the original draft, HK and KB reviewed and edited the draft, CN reviewed and edited the draft and she and AW made the supervision.

Competing interests

The authors declare that they have no competing interests.

Funding

MO|RE data is two times funded by the German Research Foundation (2014-2016 and 2021-2023).

References

1. C. Tenopir *u. a.*, „Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide“, *PLOS ONE*, Bd. 15, Nr. 3, S. e0229003, März 2020, doi: [10.1371/journal.pone.0229003](https://doi.org/10.1371/journal.pone.0229003).
2. M. C. Whitlock, „Data archiving in ecology and evolution: best practices“, *Trends in Ecology & Evolution*, Bd. 26, Nr. 2, S. 61–65, Feb. 2011, doi: [10.1016/j.tree.2010.11.006](https://doi.org/10.1016/j.tree.2010.11.006).
3. J. C. Wallis, E. Rolando, und C. L. Borgman, „If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology“, *PLOS ONE*, Bd. 8, Nr. 7, S. e67332, Juli 2013, doi: [10.1371/journal.pone.0067332](https://doi.org/10.1371/journal.pone.0067332).
4. M. Kloe, C. Niessner, A. Woll, und K. Bös, „Open Data im sportwissenschaftlichen Anwendungsfeld motorischer Tests“, *Ger J Exerc Sport Res*, Bd. 49, Nr. 4, S. 503–513, Dez. 2019, doi: [10.1007/s12662-019-00620-2](https://doi.org/10.1007/s12662-019-00620-2).

Distributed Computing and Storage Infrastructure for PUNCH4NFDI

C. Wissing¹[\[https://orcid.org/0000-0002-5090-8004\]](https://orcid.org/0000-0002-5090-8004), B. Bheemalingappa Sagar¹, M. Blank-Burian², A. Drabent³, S. Fleischer⁴, O. Freyermuth⁵, M. Giffels⁶[\[https://orcid.org/0000-0003-0193-3032\]](https://orcid.org/0000-0003-0193-3032), A. Henkel⁷, M. Hoefft³[\[https://orcid.org/0000-0001-5571-1369\]](https://orcid.org/0000-0001-5571-1369), J. Künsemöller⁸, N. Malavasi⁹, C. Manazano¹⁰, B. Roland⁶[\[https://orcid.org/0000-0003-3397-6475\]](https://orcid.org/0000-0003-3397-6475), H. Simma¹, D. Schwarz⁸[\[https://orcid.org/0000-0003-2413-0881\]](https://orcid.org/0000-0003-2413-0881), K. Schwarz¹[\[https://orcid.org/0000-0002-0800-2743\]](https://orcid.org/0000-0002-0800-2743), N. Suvvi Neelakantaia¹, L. Vomberg⁵, C. Voss¹, V. Vybornov¹⁰, M. Wigard², and S. Wozniowski¹¹

¹Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany

²Westfälische Wilhelms-Universität Münster, Germany

³Thüringer Landessternwarte, Tautenburg, Germany

⁴Helmholtzzentrum für Schwerionenforschung GmbH, Darmstadt, Germany

⁵Rheinische Friedrich-Wilhelms-Universität Bonn, Germany

⁶Karlsruher Institut für Technologie KIT, Karlsruhe, Germany

⁷Universität Mainz, Germany

⁸Universität Bielefeld, Germany

⁹Ludwig-Maximilians-Universität München, Germany

¹⁰Forschungszentrum Jülich, Jülich, Germany

¹¹Georg-August-Universität Göttingen, Germany

Abstract: The PUNCH4NFDI consortium brings together scientists from the German particle physics, hadron and nuclear physics, astronomy, and astro-particle physics communities to improve the management and (re-)use of scientific data from these interrelated communities. The PUNCH sciences have a long tradition of building large instruments that are planned, constructed and operated by international collaborations. While the large collaborations typically employ advanced tools for data management and distribution, smaller-scale experiments often suffer from very limited resources to address these aspects. One of the aims of the consortium is to evaluate and enable or adopt existing solutions. Instances of a prototype federated and distributed computing and storage infrastructure have been set up at a handful of sites in Germany. This prototype is used to gain experience in running of scientific workflows to further guide the development of the Science Data Platform, which is an overarching goal of the consortium.

Keywords: Distributed Computing, NFDI, Particle Physics, Astro-Physics, Data Management

The PUNCH4NFDI (Particle, Universe, NuClei & Hadron physics) consortium [1] brings together scientists from the German particle physics, hadron and nuclear physics, astronomy, and astro-particle physics communities to improve the management and (re-)use of scientific data from these interrelated communities. The PUNCH sciences have a long tradition of building large instruments that are planned, constructed and operated by international collaborations. Flagship examples are the particle physics experiments at the Large Hadron Collider (LHC) at CERN, which produce tens of petabytes per experiment per year, complemented by a similar amount of simulated Monte Carlo data. To store, process and analyse this data, the Worldwide LHC Computing Grid (WLCG) [2] has been established, comprising more than 160 computing centres around the world, providing more than one million CPU cores, almost 1 exabyte of disk space and around 1.5 exabytes of archival storage. By the end of the decade, the upgraded HL-LHC (High Luminosity LHC) is expected to begin taking data, requiring about an order of magnitude more CPU and storage capacity. Astronomical facilities do not yet reach the same scale but will soon catch up. Currently the radio interferometer LOFAR already relies on a federated Long Term Archive (LOFAR LTA) that holds over 50 petabytes of intermediary data products. The flagship (radio) astronomy project, the Square Kilometre Array (SKA), will have similar requirements as LHC-HL around the same time.

Although the flagship experiments are the most visible, to answer the scientific questions addressed in the PUNCH communities, many medium or small experiments are needed, each addressing a very specific scientific measurement. The amount of data produced in these experiments is typically much smaller, but issues of data management and long-term archiving are often not addressed as systematically as in the large experiments. It should be noted that the astronomy community has a long tradition of making scientific data open, including the definition of data formats and exchange protocols.

PUNCH4NFDI aims to combine the expertise of the sub-communities and to further develop the tools and to facilitate access to resources for storing, publishing and analysing data and software across all PUNCH4NFDI communities. For example, smaller experiments can benefit from the tools and methods developed for the flagship experiments. For some of the tasks, existing solutions can be adopted, while for others new solutions will be required. PUNCH4NFDI provides the forum for the knowledge transfer. Another important aspect is the ability to share existing and future computing infrastructure across the boundaries of experiments and communities for efficient use.

In this contribution, the focus is on the Compute4PUNCH and Storage4PUNCH concepts that are being developed to provide seamless and federated access to the wide variety of compute and storage systems provided by the participating communities to meet their diverse needs. Figure 1 depicts an architecture sketch. Both concepts include state-of-the-art technologies such as a token-based AAI for standardised access to compute and storage resources. For a first implementation of a PUNCH AAI, the consortium has adopted the Helmholtz AAI. Due to the diverse international participation, the AAI has to interact with existing or developing AAI structures such as the SciTokens [3] used in WLCG or EGI-Checkin and other EOSC-related implementations.

In a prototype distributed infrastructure setup, heterogeneous HPC, HTC and cloud compute resources provided by the community are dynamically and transparently integrated into a federated HTCondor-based overlay batch system using the COBaLD/TARDIS resource meta-scheduler [4]. Traditional login nodes and a JupyterHub provide entry points to the full landscape of available compute resources. Scientific software is dis-

tributed using the latest container technology and the CERN Virtual Machine File System (CVMFS) [5], which has been proven to allow replication of once centrally installed software to tens of thousands of globally distributed compute nodes via layers of HTTP Squid caches.

In Storage4PUNCH, community-supplied storage systems, mainly based on dCache [6] or XRootD technology, is federated into a common infrastructure using methods well established in the wider HEP community. Existing solutions that allow coherent management of files, which can be grouped into datasets, across multiple geographical locations will be evaluated. These systems provide methods for 'technical metadata' to manage the files, such as file sizes, checksums or file locations. Systems capable of describing the data more in terms of scientific content are also within the scope of this evaluation and prototyping. Existing caching technologies will also be evaluated for deeper integration. The combined Compute4PUNCH and Storage4PUNCH environment will enable a wide range of researchers to perform resource-intensive analysis tasks.

The current prototype spans a handful of sites in Germany that provide storage and computing resources. In addition to the purely technological evaluation of the integrated components, the first real scientific workflows are being carried out. The experience gained will help to guide further integration steps and identify the developments needed to meet the requirements of scientific applications. Over time, the prototype should incorporate more and existing data sources and thus provide a more uniform access to them. Ultimately, this distributed infrastructure should become the backbone of the Science Data Platform for the participating community, which is one of the overarching goals of the PUNCH4NFDI consortium.

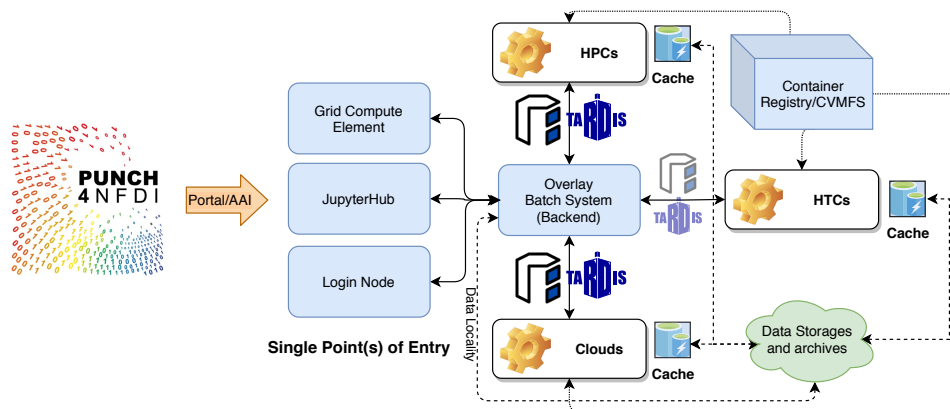


Figure 1. Architecture diagram of the distributed computing and storage PUNCH4NFDI infrastructure.

Competing interests

The authors declare that they have no competing interests.

Funding

PUNCH4NFDI is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 460248186.

Acknowledgements

The authors want to thank the entire PUNCH4NFDI consortium members for the fruitful discussions and the exchange about requirements and experiences with the prototype infrastructure.

References

- [1] The PUNCH4NFDI Consortium, *Punch4nfdi consortium proposal*, version v1 without funding tables, This is the version documenting the work plan at the proposal stage. The reduction in funding led to a re-shaping of the work programme that is documented elsewhere., Sep. 2020. DOI: [10.5281/zenodo.5722895](https://doi.org/10.5281/zenodo.5722895). [Online]. Available: <https://doi.org/10.5281/zenodo.5722895>.
- [2] I. Bird, "Computing for the large hadron collider," *Annual Review of Nuclear and Particle Science*, vol. 61, no. 1, pp. 99–118, 2011. DOI: [10.1146/annurev-nucl-102010-130059](https://doi.org/10.1146/annurev-nucl-102010-130059). eprint: <https://doi.org/10.1146/annurev-nucl-102010-130059>. [Online]. Available: <https://doi.org/10.1146/annurev-nucl-102010-130059>.
- [3] Weitzel, Derek, Bockelman, Brian, Basney, Jim, Tannenbaum, Todd, Miller, Zach, and Gaynor, Jeff, "Capability-based authorization for hep," *EPJ Web Conf.*, vol. 214, p. 04 014, 2019. DOI: [10.1051/epjconf/201921404014](https://doi.org/10.1051/epjconf/201921404014). [Online]. Available: <https://doi.org/10.1051/epjconf/201921404014>.
- [4] M. Giffels, M. Fischer, A. Haas, *et al.*, *Matterminers/tardis: The escape*, version 0.7.0, Feb. 2023. DOI: [10.5281/zenodo.7680164](https://doi.org/10.5281/zenodo.7680164). [Online]. Available: <https://doi.org/10.5281/zenodo.7680164>.
- [5] J. Blomer, P. Buncic, G. Ganis, N. Hardi, R. Meusel, and R. Popescu, "New directions in the cernvm file system," *Journal of Physics: Conference Series*, vol. 898, no. 6, p. 062 031, Oct. 2017. DOI: [10.1088/1742-6596/898/6/062031](https://doi.org/10.1088/1742-6596/898/6/062031). [Online]. Available: <https://dx.doi.org/10.1088/1742-6596/898/6/062031>.
- [6] T. Mkrtchyan, O. Adeyemi, P. Fuhrmann, *et al.*, "dCache - storage for advanced scientific use cases and beyond," *EPJ Web Conf.*, vol. 214, A. Forti, L. Betev, M. Litmaath, O. Smirnova, and P. Hristov, Eds., p. 04 042, 2019. DOI: [10.1051/epjconf/201921404042](https://doi.org/10.1051/epjconf/201921404042).

One Resource to Teach Them All

Dominik Brillhaus¹[\[https://orcid.org/0000-0001-9021-3197\]](https://orcid.org/0000-0001-9021-3197), Martin Kuhl Author²[\[https://orcid.org/0000-0002-8493-1077\]](https://orcid.org/0000-0002-8493-1077),
Cristina Martins Rodrigues³[\[https://orcid.org/0000-0002-4849-1537\]](https://orcid.org/0000-0002-4849-1537), and Andrea Schrader⁴[\[https://orcid.org/0000-0002-3879-7057\]](https://orcid.org/0000-0002-3879-7057)

¹ Data Science and Management, CEPLAS, Heinrich Heine University Düsseldorf, Germany

² Computational Systems Biology, DataPLANT, University of Kaiserslautern-Landau, Germany

³ eScience, DataPLANT, University of Freiburg, Germany

⁴ Data Science and Management, CEPLAS, University of Cologne, Germany

Abstract. Open Educational Resources (OER) allow for free access to educational materials and increase the chances of educational equity. We developed the DataPLANT OER, a teaching material resource based on the concept of annotated bricks along didactic paths. Our concept builds on a levelled approach with the brick, unit and dissemination level, is environment-agnostic and can be implemented in any desired technical framework. It balances customization and reuse and aims at one OER to teach them all.

Keywords: Open Educational Resource, OER, Teaching Material, Markdown, Metadata, YAML, front matter, didactic path

1. Introduction

Creating high-quality educational resources can be a challenging and time-consuming process. Besides providing original content, this process usually builds on the reuse of existing resources. Considering the tremendous amount of educational content available, this can be a daunting task for many educators. Teaching materials are oftentimes scattered, hard to find, and fragmented.

Following the principle of Open Educational Resources (OER) to promote easy access to educational materials and increase the chances of educational equity ([1]), a central collection of such materials should allow the community to add and improve contributions with transparent and traceable authorship and licensing of contents. We aim at one OER for all: a light-weighted, central, reusable, adaptable, open and contribution-open resource of materials for varying learning environments. Therefore, we develop the DataPLANT OER, a teaching material resource based on the concept of annotated bricks along didactic paths.

2. Three Levels to Balance Customization and Reuse

Our concept builds on a leveled approach with three levels called bricks, units and disseminations. Bricks are the smallest possible educational content snippets. Due to their compact nature, these can ubiquitously be reused. Bricks are designed for modular combination towards larger, coherent building blocks (units). A brick can be a paragraph of text or an individual slide. Units cover a complete idea, thought or topic and, thereby, are equally suited for modularization. Bricks, bricks and units or multiple units can be combined to form even larger units and eventually full-fledged disseminations.

Disseminations transport a customized didactic path specific to the learning environment of the target group or person. This can occur at a specific event or within another format like an article, knowledge base or a self-guided tutorial. Several disseminations can be combined for a more complex didactic scenario up to a didactic series built from multiple disseminations and based on a curriculum.

Teaching material can vary in format, purpose and addressed level of a learner's expertise. Each level enables maximal reuse and encourages for both, the creation of new content as well as improvement and update of existing content. This levelled approach balances the need to be able to reuse material "as-is" and the customization with own content to create materials tailored to a specific learning environment.

3. From Bricks to Disseminations with Metadata

We have designed minimal tools to assemble bricks into units or disseminations. Technically, all information required for this can be provided in one file. In its main section, this file comprises all paths to the individual files in the desired order to compile the different building blocks into units or disseminations. Didactic and other information are provided by the user in the "header" of the file called the YAML front matter section. Moreover, the user can add comments in the main section to share the idea of the didactic intention more precisely when arranging the different building blocks in addition to the descriptions in the YAML front matter. Thereby, this file is a condensation of the user's didactic path for the unit or dissemination that is linked with its implementation. It can be extended with customized style files in contrast to the default settings.

In a simple form, a brick contains the information for a single presentation slide, including contents such as a headline, a text body (e.g. with paragraphs, bullet points) and the reference to an image to be presented. Bricks are written in markdown ([2]) which allows to add metadata in the YAML front matter section, which is not displayed in the final outcome.

Inspired by recommendations of [3] and [4], we use the YAML front matter section for standardized terms covering didactic (educational), legal and technical metadata in bricks, units and disseminations and all files compiled thereof. Virtually, any kind of metadata can directly be associated with the teaching material. Didactic metadata attributes include, e.g., learning outcomes, skills, requirements, target audience and teaching mode. The legal section focuses on material authorship and licenses and allows to define terms of reuse "in both directions", i.e. what is the source to a material that was reused while creating content (like a brick) and under what conditions can the created content itself be reused. The technical section aligns with the chosen technical implementation like the interpreter of the markdown content and the associated styling or formatting to render the outcome. The types of attributes differ between the three levels and aggregate from brick over unit to dissemination.

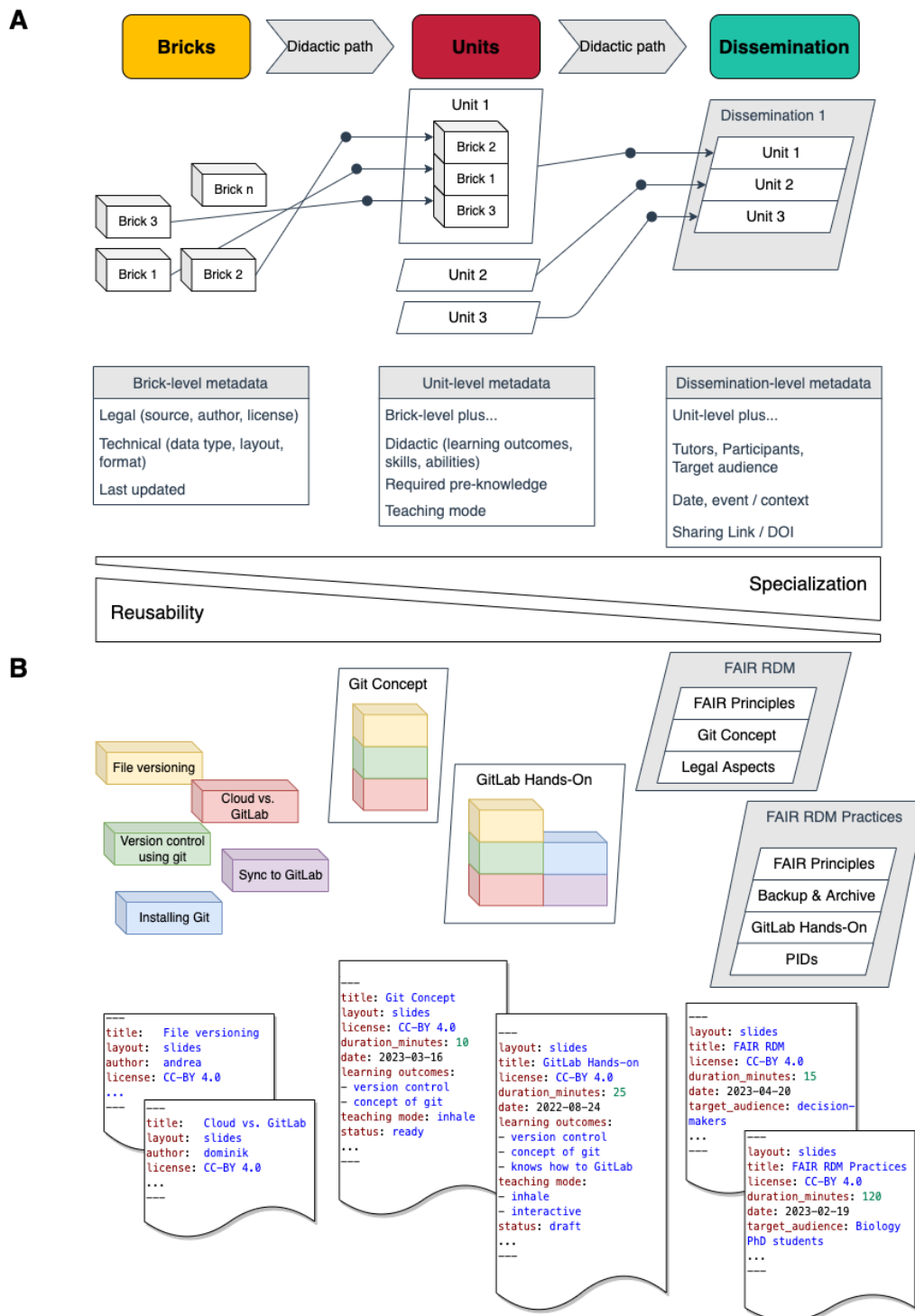


Figure 1: A) Modular concept for the DataPLANT OER. The educator can assemble any number of bricks into a unit. Together with other units (preassembled or customized), the educator can then compile the teaching content for a dissemination. This concept offers the perfect blend of content reuse and customization for distinct learning environments (courses, workshops, etc.).

B) Example “version control using Git” to elaborate our leveled concept. One could design individual bricks to elaborate the benefits of file versioning, how Git may be helpful for this purpose, how Git platforms like GitHub and GitLab overlap with or are different from other cloud services, etc. These bricks could now be compiled in an independent unit, teaching Git version control only on a conceptual level without much technical depth. This unit could be taught to an audience with little technical or coding background. The same bricks

could be combined with more technical details on how to install Git, how to Git version control local files and share them via Git-backed platforms to compile a unit as the basis for a tutorial, demo and workshop – three different formats – to transfer hands-on knowledge to an audience that would actually employ Git. Both units could eventually be assembled with other units to design a dissemination for a larger context such as a course about “FAIR RDM Practices” targeted at PhD students.

4. Brief Notes on the Technical Implementation

Our basic levelled concept is environment-agnostic and could be implemented in any desired technical framework. For teaching materials collected in DataPLANT, we have decided to create bricks (and consequently units and disseminations) in form of “markdown”-formatted text files and collect these together with images in a shared GitHub repository (see example at [5]) aiming at an OER.

Using the YAML front matter, different markdown interpreters can render the provided content to a defined output like the Marp presentation ecosystem ([6]) or reveal.js ([7]) for markdown-based slides, static website generators such as Fornax ([8]) or pandoc ([9]), which allows to convert into virtually any relevant document file type. As such, the same input (markdown text) can be rendered to different outputs (slides, articles). Designed as pure text files, markdown files represent an open file format and work well with text-based version control such as Git ([10]). Also, the schematized YAML front matter allows to programmatically read out the metadata attributes to build a database to facilitate findability and reuse of teaching materials.

5. Conclusion & Outlook

Our envisioned concept offers great chances for spreading the spirit for community endeavors and driving the use of OERs. Despite the learning curve at the beginning (how to write markdowns, how to assemble units and disseminations, and maybe even how to use Git), users can benefit immensely from the balance between reusability and specialization. While DataPLANT data stewards support content creators, we also work on easier in(-and-out) routes, interoperability as well as user friendly and light maintenance options for all user flavors.

Data availability statement

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

We acknowledge the support of DataPLANT, funded by the German Research Foundation (DFG) within the framework of the NFDI – project number: 442077441 and CEPLAS, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within Germany’s Excellence Strategy – EXC-2048/1 – project ID 390686111.

References

1. UNESCO, "The 2019 UNESCO Recommendation on Open Educational Resources (OER): supporting universal access to information through quality open learning materials". [Online]. UNESDOC Digital Library, Catalog Number 0000383205, 2022. Accessed: Apr. 21, 2023. Available: <https://unesdoc.unesco.org/ark:/48223/pf0000383205.locale=en>
2. J. Gruber, Markdown (2004). [Online]. Available: <https://daringfireball.net/projects/markdown/>
3. L. Garcia et al., "Ten simple rules for making training materials FAIR", *PLoS Comput Biol*, vol. 16, no. 5, e1007854, May 2020. [Online]. Available: <https://doi.org/10.1371/journal.pcbi.1007854>
4. B. Batut et al., "Design and plan session, course, materials (Galaxy Training Materials)." Accessed: Apr. 21, 2023. [Online]. Available: <https://training.galaxyproject.org/training-material/topics/contributing/tutorials/design/tutorial.html>
5. D. Brilhaus, M. Kuhl, C. Martins Rodrigues and A. Schrader "Teaching material - Concept" [Online]. Available: <https://Github.com/nfdi4plants/teaching-materials-concept>
6. *Marp*. (2023). Y. Hattori, @marp-team. [Online]. Available: <https://Github.com/marp-team/marp/>
7. *Reveal.js*. (2020). Hakim El Hattab (@hakimel). [Online]. Available: <https://revealjs.com/markdown/>
8. *Fornax*. (2020). Ionide. [Online]. Available: <https://Github.com/ionide/Fornax>
9. *Pandoc*. (2023). J. MacFarlane. [Online]. Available: <https://pandoc.org/>
10. *Git*. (2023). Software Freedom Conservancy. [Online]. Available: <https://Git-scm.com/>

Transparency and Involvement of the Energy-Related Industry in a Data Sharing Platform

Zhiyu Pan¹, Gonca Gürses-Tran², Christina Speck³, Patrick Jaquart³, Michael Niebisch⁴, and Antonello Monti^{1,2}

¹RWTH Aachen University, Germany

²Fraunhofer FIT, Germany

³Karlsruhe Institute of Technology, Germany

⁴Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Abstract: The integration of renewable energy sources, the decentralization of the energy system, and the increasing digitization of energy-related processes require the integration of a wide range of energy-related data. In this context, a data sharing platform can serve as a hub for exchanging energy-related data and developing innovative solutions to improve the efficiency and sustainability of the energy system. However, especially because of the involvement of the energy-related industry in such a platform poses several challenges related to data protection, intellectual property, and business interests. This paper presents a framework for ensuring transparency and involvement of the energy-related industry in a data sharing platform, based on the FAIR data principles and a co-creation approach involving industry partners.

Keywords: Data sharing platform, FAIR principles, energy industry

Introduction

The transformation of the energy system towards a low-carbon and decentralized model requires the integration of a wide range of energy-related data, including electricity production and consumption, weather conditions, energy storage, and grid infrastructure. However, much of this data is currently dispersed across different stakeholders, such as utilities, grid operators, regulators, and consumers, and is subject to various legal, technical, and economic barriers to sharing. To overcome these challenges, a data sharing platform can provide a common space for collecting, processing, and sharing energy-related data among different actors, thus enabling the development of new services, applications, and business models based on data-driven insights.

However, the involvement of the energy-related industry in a data sharing platform requires a careful balance between the interests of different stakeholders, such as data providers, data users, and platform operators. On the one hand, data providers may have concerns about data protection, intellectual property, and privacy, as well as about the potential competitive advantage that their data may provide to other actors. On the other hand, data users may have requirements for data quality, interoperability, and

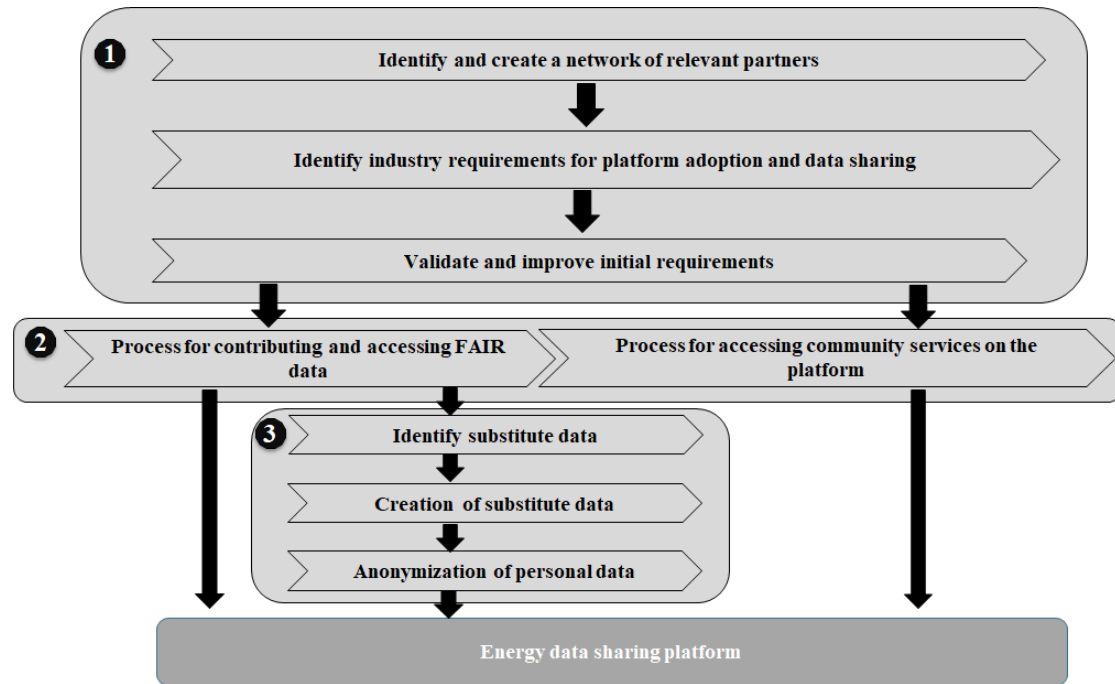


Figure 1. Framework for ensuring transparency and involvement of the energy-related industry in a data sharing platform

availability, as well as for fair and transparent conditions for accessing and using the data. In addition, platform operators need to ensure the sustainability, scalability, and security of the platform, while also fostering a collaborative and inclusive environment for data sharing and innovation. This is crucial for establishing trust among industry partners, promoting long-term engagement, and ensuring that the platform is capable of meeting the evolving needs of the energy sector. To address these challenges, this paper proposes a framework for ensuring transparency and involvement of the energy-related industry in a data sharing platform, based on the FAIR data principles [1].

1 Methodology

The proposed framework in Figure 1 consists of three main parts: (1) the definition of technical and organizational requirements for data sharing, (2) the involvement of industry partners in the co-creation of the platform, and (3) the collection and creation of substitute data.

In the first part, the relevant partners for the energy domain should be identified. The technical and organizational requirements for data sharing are based on the FAIR data principles, which provide guidelines for making data Findable, Accessible, Interoperable, and Reusable. This includes the use of standardized data formats, metadata, and vocabularies, as well as the provision of appropriate documentation, licenses, and identifiers. In addition, the platform should support data quality control, data enrichment, and data integration services, to ensure that the data is relevant, accurate, and consistent across different sources. To facilitate this process, the platform provides collaborative and participatory tools, such as forums, workshops, and hackathons, that allows industry partners to exchange ideas and feedback.

The involvement of industry partners in the co-creation of the platform is essential to ensure that the platform meets the needs and expectations of the energy-related industry. This includes the identification of data sources, data use cases, and data sharing

agreements, as well as the development of new services, applications, and business models based on the data. Besides data, the industry requirement for accessing and using the related services is also part of the framework.

The last part of the framework is the collection and creation of substitute data. The industry partners may not always be able to provide complete data sets due to privacy concerns or other reasons. Therefore, synthetic data will be created where necessary to ensure that the platform has the data required for analysis. Tools for the anonymization of personal data are integrated to ensure that the privacy of individuals is protected.

2 Conclusion

In conclusion, the framework developed addresses the need for transparency and involvement of the energy-related industry by creating a collaborative environment where industry partners can contribute and access FAIR data, access community services, and provide feedback for continuous improvement. Additionally, the framework addresses the issue of missing or personal data through the development of tools for the creation of synthetic data and the anonymization of personal data. By incorporating industry needs and concerns, the framework facilitates collaboration between industry partners and researchers, resulting in more effective and efficient energy systems.

Data availability statement

Not applicable.

Underlying and related material

Not applicable.

Author contributions

Conceptualization, Z.P., G.G., C.S., P.J., M.N.; methodology, Z.P. writing—original draft preparation, Z.P.; writing—review and editing, Z.P., G.G., C.S., P.J., M.N.; supervision, A.M..

Competing interests

The authors declare that they have no competing interests.

Funding

The authors would like to thank the German Federal Government, the German State Governments, and the Joint Science Conference (GWK) for their funding and support as part of the NFDI4Energy consortium. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 501865131.

References

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

WissKI Viewer

Casual Access for WissKI Data Sets

Tom Wiesing¹ [<https://orcid.org/0009-0002-7392-0556>]

¹FAU Erlangen-Nürnberg, Germany

Abstract: WissKI is a software which allows researchers to record data about objects of the cultural heritage in a graph database backed by a formal ontology. It acts as a database for researchers to store their results via a web interface, and nearly automatically makes data FAIR, linked and open. To install the WissKI software a nontrivial amount of effort is required and typically requires help from a system administrator. To make it easier to access WissKI and data stored within this paper introduces the WissKI Viewer, which enables researchers to directly inspect a WissKI Backup on their own computer, without additional help.

Keywords: WissKI, FAIR, Linked Open Data

FAIR [1] is an acronym defined as Findable, Accessible, Interoperable and Reusable. It is a desired quality of research data, and is self-explanatory. Findable means that data can be identified, and is well-described with appropriate metadata. Accessible means that data can be accessed, and does not require custom authentication or authorization schemes. Interoperable means that data uses common data formats, and is machine-readable. Reusable means that data can be reused, and is not limited to the original purpose it was created or collected for.

Linked Open Data is an orthogonal concept to FAIR meaning data both open and linked. Open means being available under an Open Source license. Linked means being interlinked and accessible using semantic queries.

A Triplestore, also known as a graph database, consists of (subject, predicate, object) triples describing objects and their properties. Triples correspond to edges in a labeled, directed graph. The subject is the source node, the object the sink node, and the predicate the label.

Most common triplestore implementations are based on RDF [2], and can be queried using SparQL [3]. SparQL is ideal for enabling linked data, as it provides a formal query language. It allows provides facilities more making data linked and open using so-called federated queries.

Drupal [4] is an extensible open source content-management system written in a programming language called PHP. It is available via a web interface, and enables users to create and manage content on their website. The backend of Drupal is powered by an SQL database.

WissKI [5], [6] is a system which allows researchers to record data about objects of the cultural heritage in a graph database backed by a formal ontology. The WissKI interface provides three main functionalities.

- An administrative interface where a formal ontology can be defined and edited,
- An edit interface where database entries can be added and edited without the need to entirely understand the formal ontology, and
- A public-facing browsing interface that allows browsing the database.

For any particular research project, the WissKI software first has to be installed on a server accessible for the desired users. It then has to be extended in order to fit the projects needs. Such a WissKI-based system consists of several components which can be seen in Figure 1. They are implemented as a set of extensions (called modules) to the content management system Drupal and accessible to users via a web interface.

Drupal Core The Drupal Core represents an installation of the Drupal content management system. It handles authentication and manages different display options for content.

SQL Database Required by Drupal in order to store authentication and presentation configuration data. It also acts as a cache.

WissKI modules The WissKI modules run inside of drupal and implement the main functionality of WissKI.

Triplestore Data entered by WissKI users – the research result – is stored inside a triplestore.

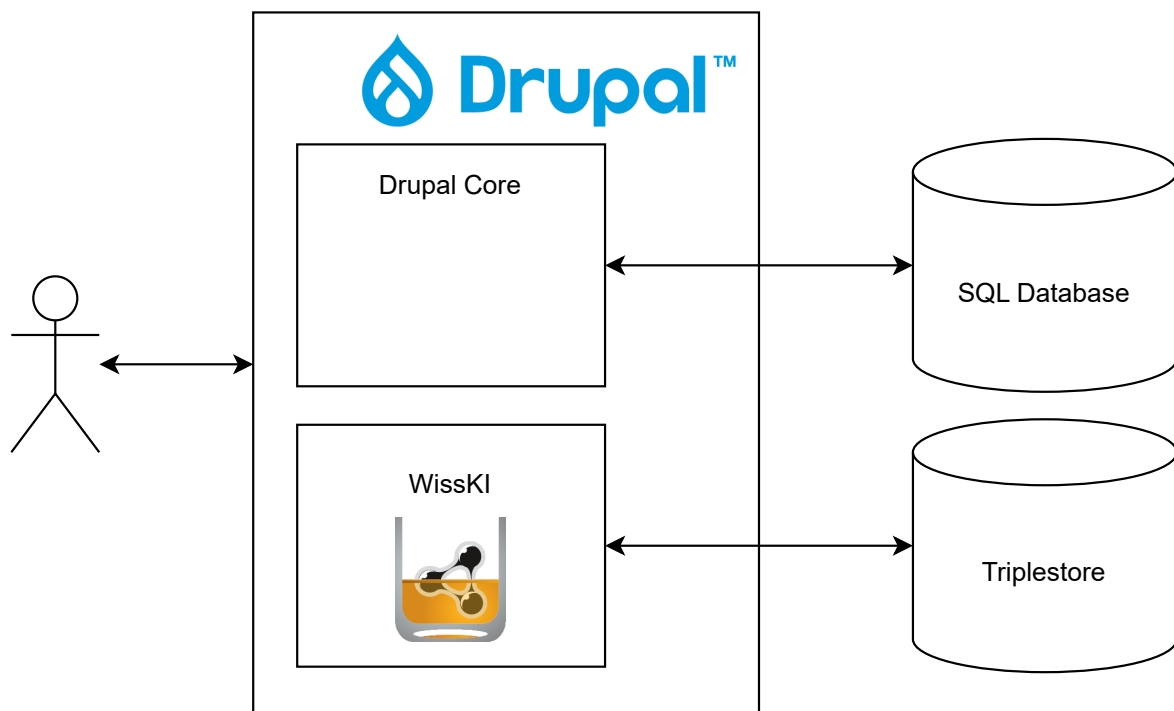


Figure 1. Overview of a WissKI-based system

In order for a WissKI-based system to function properly, each of these components requires separate configuration. This makes it difficult to set up and maintain - typically a system administrator with significant expertise is required. Furthermore, once a research project has ended and funding has run out it quickly ends up in an unus-

able state or is shutdown entirely. Only the research result – the data in the triplestore – survives this shutdown. Information regarding how data in the database should be interpreted by users - such as formatting - does not survive.

A long term goal should be to avoid such a shutdown by improving the maintainability and usability of the WissKI software itself. But this alone is insufficient to fully address the problem – providing a shared interface to view, edit and manage a dataset has inherit complexity and comes with maintenance implications. Instead the WissKI Viewer takes a different approach - it is a software that only aims at making WissKI datasets viewable. This is sufficient to ensure that the research result remains viable after a project ends.

The WissKI viewer runs directly on a researchers computer and does not require any systems knowledge in order to function. It directly provides the researcher with an interface to view any database entries created in the originating system. The WissKI Viewer is written in go and available at [7]. It is distributed as a standalone application, and a screenshot of the viewer interface can be seen in Figure 2.

<http://kirmes.wisski.agfd.fau.de/#5f15b12cd5fb1>

[Bundles] > [Bundle Werke] > [Entity <http://kirmes.wisski.agfd.fau.de/#5f15b12cd5fb1>]

Fields

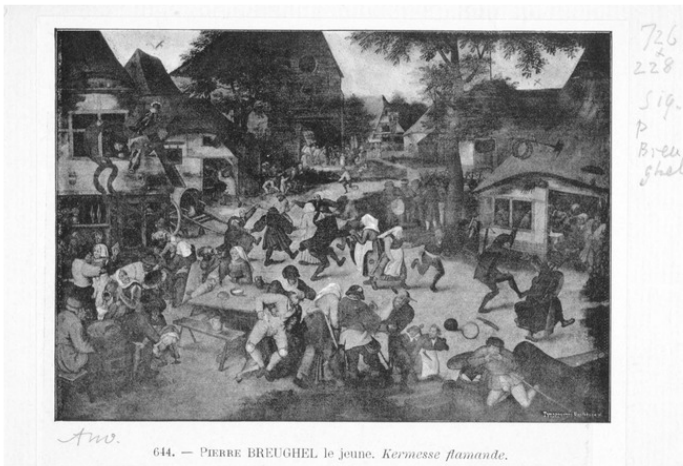
Name	Type	Count	Value(s)
Show Entity Triples			
Werktitel/Title	string	1 (Cardinality 1)	Dorpskermis op het feest van de H. Joris
Abbildungs-ID/Image-ID	string	1 (Cardinality -1)	8f142990-0a89-8332-a28b-25375773d9c6
Abbildung/Image	image	1 (Cardinality -1)	
Kommentar/Comment	text_long	1 (Cardinality 1)	Diese Version eventuell identisch mit dem Gemälde , das Max Friedländer in Antwerpen vermutete. Es ist unten links mit P. BREUGHEL signiert. Es befindet sich laut Eintrag in der Datenbank des RKD heute im Koninklijk Museum voor Schone Kunsten Antwerpen, (inv./cat.nr 644), als Schenkung von A. Michiels. Eine weitere Version von Pieter Bruegel dem Jüngeren wurde 2004 bei Sotheby's versteigert.

Figure 2. A database entry about a painting displayed in the WissKI Viewer.

The viewer takes as input the surviving graph data from the triplestore as well as a configuration file containing the ontology used. Upon startup, it first loads the graph data into memory, and then repeatedly scans it according to the ontology used to recover the original entries. Depending on the exact size of the dataset, this process can

take a few seconds (for datasets consisting of a few thousand triples) to a few minutes (for datasets consisting of several million triples).

Compared to WissKI itself, the viewer does not rely on any external databases. This makes it ideal to ensure that WissKI-produced datasets remain accessible after the original system that produced them has been retired.

Author contributions

We contributed by developing the WissKI Viewer based on the pre-existing WissKI system.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The author wishes to acknowledge the help of the WissKI community, and in particular Mark Fichtner, for help during the development of the WissKI Viewer.

References

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, 2016. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
- [2] World Wide Web Consortium (W3C), Ed. "Resource description framework (RDF)." (), [Online]. Available: <http://www.w3.org/RDF/> (visited on 10/22/2009).
- [3] "SPARQL 1.1 overview," World Wide Web Consortium (W3C), W3C Recommendation, Mar. 23, 2013. [Online]. Available: <https://www.w3.org/TR/sparql11-overview/>.
- [4] Drupal Association, *Drupal - Open Source CMS*, <https://www.drupal.org/>, [Online; accessed 20-April-2023], 2023.
- [5] G. Hohmann and B. Schiemann, "An Ontology-Based Communication System for Cultural Heritage. Approach and Progress of the WissKI Project," in *Scientific Computing and Cultural Heritage. Contributions in Computational Humanities*. (Contributions in Mathematical and Computational Science), H. G. Bock, W. Jäger, and M. Winckler, Eds., Contributions in Mathematical and Computational Science. Berlin: Springer, 2013, vol. 3, pp. 127–135.
- [6] G. Goerz, "WissKI: Semantische Annotation, Wissensverarbeitung und Wissenschaftskommunikation in einer virtuellen Forschungsumgebung," Jun. 2011.
- [7] Tom Wiesing, *hangover - A WissKI Viewer*, <https://github.com/FAU-CDI/drincw#wisski-viewer--exporter>, Accessed on April 19, 2023, 2023.

Pathways Between National and European Research Infrastructures: A Humanities' perspective

Nanette Reißler-Pipka⁶[\[https://orcid.org/0000-0002-0719-9003\]](https://orcid.org/0000-0002-0719-9003), Regine Stein²[\[https://orcid.org/0000-0003-3406-5104\]](https://orcid.org/0000-0003-3406-5104),
Laure Barbot³[\[https://orcid.org/0000-0002-6008-7959\]](https://orcid.org/0000-0002-6008-7959), Sally Chambers^{3,4}[\[https://orcid.org/0000-0002-2430-475X\]](https://orcid.org/0000-0002-2430-475X), Toma
Tasovac^{3,5}[\[https://orcid.org/0000-0002-3919-993X\]](https://orcid.org/0000-0002-3919-993X) and Philipp Wieder¹[\[https://orcid.org/0000-0002-6992-1866\]](https://orcid.org/0000-0002-6992-1866)

¹ Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, Germany

² Göttingen State and University Library, Germany

³ DARIAH ERIC, France

⁴ Ghent Centre for Digital Humanities, Ghent University, Belgium

⁵ Belgrade Center for Digital Humanities, Serbia

⁶Max Weber Foundation, Germany

Abstract. In this poster we will discuss challenges, requirements and good practices for creating synergies and inter-connections between national and international research and research data infrastructures from the perspective of the humanities. Stakeholders of the NFDI consortium Text+, of the German national node of DARIAH, of the SSH Open Cluster and of related EOSC projects are presenting benefits and outcomes of these structures and their linkages for the actual research data management and for the researchers. How are research data repositories connected and researchers guided to all the various service offers? With an example about the tool registry in the SSH Open Marketplace reused by Text+ and connected to EOSC, we illustrate the overall picture of these infrastructures and their impact in practice.

Keywords: Research Data Management, Humanities, Text+, DARIAH, EOSC, SSHOC, Data Spaces, Tool Registry, NFDI

Connecting the strategies and service offerings of the various research and research data infrastructures between the national and international level is a key challenge in current developments. The humanities in Germany have a long record in building digital research infrastructures in the context of the ESFRI roadmap [1]. Germany has actively contributed to establishing and developing the CLARIN [2] and DARIAH [3] ERICs and is supporting OPERAS [4] on its way to become an ERIC [5]. With CLARIN's focus on digital research infrastructure for language resources, DARIAH's wider perspective on the arts and humanities and OPERAS' pushing forward open scholarly communication, a number of large and interconnected research communities are being addressed and their German national nodes are connected under the umbrella of the Association for Research Infrastructures in the Humanities and Cultural Studies [6].

Stakeholders of these infrastructures are strongly involved in the humanities' consortia of the NFDI, in particular in Text+ [7, 8], and thereby build an institutional bridge between the German NFDI and the European level. They are continuously contributing to the sustainable organisation of the Social Sciences and Humanities Open Cluster (SSHOC) [9] and its marketplace as one of the five science clusters in EOSC [10] and participate in EOSC-related projects such as EOSC Future [11] or FAIRCORE4EOSC [12]. However, what are the benefits

and outcomes of these structures and their linkages for actual research data management and for researchers? How are research data repositories connected and researchers guided to all the various service offers?

In this paper we illustrate the overall picture of these infrastructures and their impact through a practical example:

Reusability of registries: Tool registry in Text+ and in SSH Open Marketplace

The NFDI builds on valuable previous work and promises a long-term, reliable and sustainable infrastructure for its research communities. A typical user and community requirement in the NFDI is an easily understandable overview of the offerings. In addition to data, consultancy and training resources, tools and services should be maintained, registered, curated and displayed. Text+ will rely on the existing tool registry in the SSH Open Marketplace and combine it with its own specific search functionalities. The SSH Open Marketplace is one of the key exploitable results of the SSHOC project [13, 14, 15]. Two features of the tool registry are essential: 1. Open collaboration in the marketplace using EOSC-compatible / DARIAH-AAI single-sign-on login paired with an editorial board which reviews and approves new entries and changes. 2. Technical readiness to be integrated into other platforms via API and a trusted hosting in DARIAH-EU. The SSH Open Marketplace has already been registered as a service in the EOSC catalogue and aims to make relevant tools and services listed in the SSH Open Marketplace also findable and visible in the EOSC Marketplace. In this way, the tools and services from Text+ become integrated within EOSC.

Through this example, challenges, requirements and good practices for creating synergies and inter-connections between national and international research and research data infrastructures will be shown from the perspective of the humanities.

Data availability statement

Not applicable.

Underlying and related material

Not applicable.

Author contributions

Nanette Rißler-Pipka (Writing - original draft); Regine Stein (Writing - original draft); Laure Barbot (Writing - review & editing); Sally Chambers (Writing - review & editing); Toma Tasovac (Writing - review & editing); Philipp Wieder (Writing - review & editing)

Competing interests

The authors declare that they have no competing interests.

Funding

Text+ receives funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 460033370.

The EOSC Future project is co-funded by the European Union Horizon Programme call INFRAEOSC-03-2020 - Grant Agreement Number 101017536.

Acknowledgement

Not applicable.

References

1. ESFRI Roadmap, Website. <https://roadmap2021.esfri.eu/> (accessed 26/04/2023)
2. CLARIN, Website. <https://www.clarin.eu/> (accessed 26/04/2023)
3. DARIAH, Website. <https://www.dariah.eu/> (accessed 26/04/2023)
4. OPERAS, Website. <https://operas-eu.org/> (accessed 26/04/2023)
5. ERIC, Website. European Research Infrastructure Consortium https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/european-research-infrastructures/eric_en (accessed 26/04/2023)
6. Association for Research Infrastructures in the Humanities and Cultural Studies/ (Ver- ein Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen e.V., GKFI, Website. <http://forschungsinfrastrukturen.de/doku.php/start> (accessed 26/04/2023)
7. NFDI consortium Text+, Website. <https://www.text-plus.org/en/home/> (accessed 26/04/2023)
8. Hinrichs, Erhard, Alexander Geyken, Peter Leinen, Andreas Speer, Regine Stein, Jonathan Blumtritt, Luise Borek, et al. (2022). 'Text+: Language- and Text-Based Research Data Infrastructure'. <https://doi.org/10.5281/zenodo.6452002>
9. SSH Open Cluster, Website. <https://www.sshopencloud.eu/news/sshoc-ssh-open-cluster> (accessed 26/04/2023)
10. Lamanna, Giovanni, Ian Bird, Andreas Petzold, Ari Asmi, Magdalena Brus, Niklas Blomberg, Michael Räß, Rudolf Dimper, Andrew Gotz, & Ron Dekker. (2021). ESFRI Science Clusters Position Statement on Expectations and Long-Term Commitment in Open Science (1.02). Zenodo. <https://doi.org/10.5281/zenodo.4892245>
11. EOSC Future, Website. <https://eoscfuture.eu/about/> (accessed 26/04/2023)
12. FAIRCORE4EOSC, Website. <https://faircore4eosc.eu/about> (accessed 26/04/2023)
13. Barbot, Laure, Yoan Moranville, Frank Fischer, Clara Petitfils, Matej Ďurčo, Klaus Illmayer, Tomasz Parkoła, Philipp Wieder, & Sotiris Karampatakis. (2019). SSHOC D7.1 System Specification - SSH Open Marketplace (1.0). Zenodo. <https://doi.org/10.5281/zenodo.3547649>
14. SSHOC. (2022). SSHOC Legacy booklet. Zenodo. <https://doi.org/10.5281/zenodo.6394462>
15. Ďurčo, Matej, Laure Barbot, Klaus Illmayer, Sotiris Karampatakis, Frank Fischer, Yo- ann Moranville, Joshua Tetteh Ocansey, Stefan Probst, Michał Kozak, Stefan Bud- denbohm, & Seung-Bin Yim. (2021). 7.2 Marketplace – Implementation (v1.0). Zenodo. <https://doi.org/10.5281/zenodo.5749465>

Architectural Design of BERD Information Portal

Ahmed Saleh¹[\[https://orcid.org/0000-0003-0581-4655\]](https://orcid.org/0000-0003-0581-4655), and Klaus
Tochtermann^{1,2}[\[https://orcid.org/0000-0003-2471-2697\]](https://orcid.org/0000-0003-2471-2697)

¹ZBW – Leibniz Information Centre for Economics, Kiel/Hamburg, Germany

²Kiel University, Germany

Abstract: The management of Business, Economic, and Related Data can be a complex task involving various types of data and diverse research needs. BERD@NFDI develops a powerful platform to collect, process, analyze, and preserve such data in one place. The technical architecture of the BERD platform consists of several components, each responsible for specific functionalities that work together to deliver content and services to users. At a high level, BERD infrastructure is designed to align with the recommendations of the EOSC Interoperability Framework¹, which includes several layers beyond the technical layer, such as organizational interoperability, legal interoperability, and semantic interoperability. However, this document focuses solely on the technical interoperability of the BERD architecture.

At the base layer of the BERD infrastructure, there is a physical infrastructure that includes servers, storage devices, and networking equipment, providing computing power and storage space. On top of this, there is a software infrastructure that includes the operating system, web server, and application server, providing the necessary software components to run the platform. The data layer includes the database server and any other data storage systems used by the platform to store and manage data, such as user account information and data marketplace records. The presentation layer is responsible for rendering web pages and providing an intuitive and engaging user experience. The infrastructure that hosts the BERD platform is provided by a commercial cloud provider, and it consists of four main servers: the test server, the production server, the services server, and the mockups server. The BERD Platform is built on top of InvenioRDM², an open-source research data management platform. The services used by the platform include OpenSearch³, PostgreSQL⁴, Redis⁵, and RabbitMQ⁶. OpenSearch provides a distributed, multitenant-capable full-text search engine. PostgreSQL provides a database management system (RDBMS). Redis handles the cache of the BERD Platform, for example, storing user sessions and caching rendered pages. RabbitMQ holds the tasks for Invenio workers to execute. Invenio workers are background processes or threads that execute various tasks in the Invenio software stack, such as processing and indexing metadata records, handling user

¹<https://eosc-portal.eu/eosc-interoperability-framework>

²<https://inveniordm.web.cern.ch>

³<https://opensearch.org>

⁴<https://www.postgresql.org>

⁵<https://redis.com>

⁶<https://www.rabbitmq.com>

requests, and performing maintenance tasks. All of these services, with the exception of PostgreSQL, are deployed on the servers using Gitlab-CI and docker containers. Postgres is provided as a service by the cloud provider.

The architecture of the platform consists of four main components: crawlers, a data model, an ingestion service, and a machine learning component. Crawlers are responsible for harvesting data from online sources such as digital libraries, repositories, social media platforms, and other online repositories. The data collected by crawlers is then processed by a data model that aligns the data format and cleans it up as needed. The ingestion service is responsible for inputting the processed data into a database, where it can be queried and analyzed. Finally, a machine learning component can be used to enrich the data with additional insights and patterns.

Overall, the technical architecture of the BERD platform is designed to be scalable, reliable, and secure, while also providing a high-performance and user-friendly experience for visitors.

Keywords: BERD Infrastructure, Crawlers, Search Engine

Because Data Shall Grow (and so shall we)

Steps Towards a Cultural Change for Sharing Research Data

Julia Rakers¹, Bernhard Miller²[\[https://orcid.org/0000-0002-4385-7245\]](https://orcid.org/0000-0002-4385-7245), Julia Mohrbacher³[\[https://orcid.org/0009-0005-9732-1285\]](https://orcid.org/0009-0005-9732-1285), Daniel Nüst⁴[\[https://orcid.org/0000-0002-0024-5046\]](https://orcid.org/0000-0002-0024-5046), Torsten Schrade⁵[\[https://orcid.org/0000-0002-0953-2818\]](https://orcid.org/0000-0002-0953-2818), Jörg Seegert⁴[\[https://orcid.org/0000-0001-9357-2830\]](https://orcid.org/0000-0001-9357-2830), Christian Vater⁵[\[https://orcid.org/0000-0003-1367-8489\]](https://orcid.org/0000-0003-1367-8489), Cord Wiljes⁶[\[https://orcid.org/0000-0003-2528-5391\]](https://orcid.org/0000-0003-2528-5391), Holger Simon⁷[\[https://orcid.org/0000-0001-7352-4006\]](https://orcid.org/0000-0001-7352-4006)

¹Duisburg-Essen University

²GESIS - Leibniz-Institute for the Social Sciences, Mannheim and Cologne

³Albert-Ludwigs-Universität Freiburg

⁴Technische Universität Dresden

⁵Akademie der Wissenschaften und der Literatur | Mainz – Digitale Akademie

⁶Nationale Forschungsdateninfrastruktur (NFDI) e.V.: Karlsruhe and Bielefeld University

⁷Pausanio GmbH & Co. KG

Abstract. Research data are a valuable asset of their own and individual researchers as well as the research community as a whole can benefit from data sharing practices. These benefits include but are not limited to higher data quality or the more efficient use of resources. Despite these potential gains, data sharing is not widespread yet and processes of cultural change are needed to reap the benefits of data sharing. A transition to FAIR and open data sharing can only be sustainable if it permeates all aspects of academia. Therefore, this transition takes time and must start as early as possible. In 2023, the NFDI is complete in its first incarnation with 26 funded consortia and the funded association of consortia, Base4NFDI, and can function as a platform for discussion and collaboration around cultural change. The NFDI provides a network that extends beyond individual research bubbles in the name of common interests and facilitates cultural change processes towards data sharing. We identified four central clusters of interest including 1. policies, strategies, and funding; 2. communities, workshops, and multipliers; 3. publications, and 4. collaboration, communication, and error cultures. To deepen our understanding of ongoing cultural change, a scheme for collecting use cases has been developed and an analysis of these use cases will be published to summarize learnings for the NFDI on how to encourage cultural change.

Keywords: Cultural Change, Data Sharing, FAIR Principles, NFDI

1. Research Data as a Resource for an Innovation Driven Society

Data seen as a resource is of a completely new type: unlike gold, shared data is not lost; its worth is increasing by connecting it to other resources and allowing for new uses. Unlike oil, data cannot be depleted by sharing. A society that commits to sharing research data openly is choosing a beneficial strategy.

1.1 The NFDI's Role and Purpose in this Process

Cultural change in research is primarily about designing and enabling collective and collaborative practices, e.g., quality standards or evaluation and funding criteria, that are followed and embraced by entire communities. It is grounded in a set of shared values and a common understanding that FAIRness [FAIR] and openness are more relevant than currently prevailing means to define impact and value. Cultural change needs to involve all parts of the common infrastructure, so that systems designs – hardware, software, standards, rules, guidelines – furthers the new collaborative and collective culture. This new culture has to spread widely to be sustainable and effective and must transcend the borders of institutions and disciplines. The NFDI consortia are designed to develop and promote a culture of data sharing [BLV] but jointly they can support a more far-reaching cultural change than individual research communities. The NFDI provides a common ground for discussion and negotiation, is a stable base with sustainable infrastructures, and has a broad reach to sustain cultural change. In 2023, the NFDI is complete in its first incarnation with 26 funded consortia and the funded association of consortia, Base4NFDI, and therefore the basis for driving transdisciplinary cultural change processes within the context of the NFDI is given.

1.2 Data as Humus – the CC-BY-US Series

But what should we compare data to, if not to “Gold” or “Oil”? An interesting metaphor could be “humus”. In a shared information infrastructure, a project can grow on FAIRly shared data. “Data” would thus be seen and framed as a force or an element of a shared ecology. The “CC-BY-US” series of collaborative workshops on cultural change in the NFDI will thereby promote ecological growth. Nothing needs to be thrown away in the collective information infrastructure. Instead, the research data of today will be the humus for the ideas and innovation of tomorrow. The individual contribution is made with a sense of purpose towards the future generations of “growers” of research, restraining oneself from the need of individual and immediate benefits towards a more sustainable NFDI.

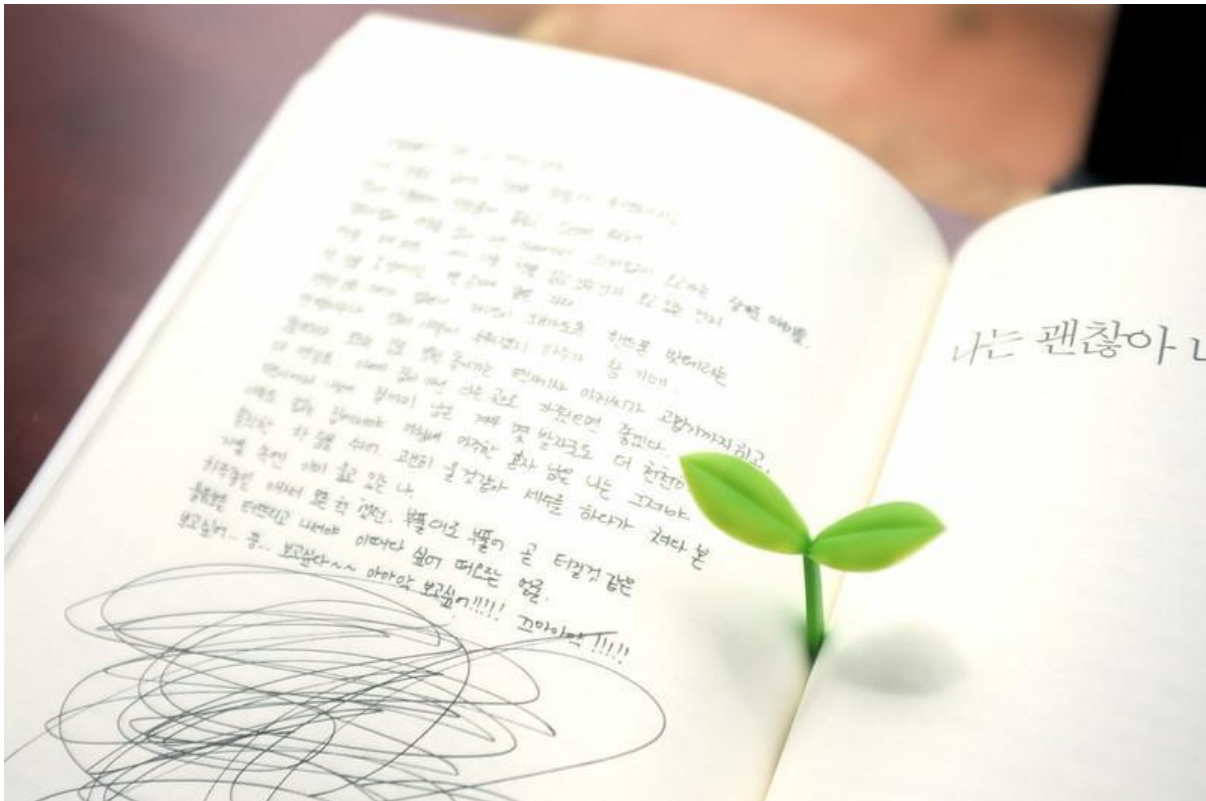


Figure 1: Data is humus. It helps research grow. Source: flickr.com, k.o.u - [Sprout!](#) (CC-BY 2.0)

2. Fields of Action: Fertilizing Growth of FAIR Data

In a first workshop “CC-BY-US: Cultural change in sharing research data with and through the NFDI” [CCBYUS1] participants from across the NFDI identified four main clusters of interest and focused on explicating some basic opportunities for collaboration and examples of best practice. The event also identified old habits to overcome and general threats to the endeavor. These clusters are:

- Policies, strategies, and funding (e.g., data stewardship [TUDelft]; including limitations [DSSS] [TW])
- Communities (e.g., pledges [N4E], considering target groups), multipliers, workshops, and surveys (e.g., [FDM])
- publications (e.g., contribution statements [CRedit])
- collaboration, communication, and error culture

A second workshop “CC-BY-US 2: How can cultural change promote the use of infrastructures; how do infrastructures promote cultural change?” [CCBYUS2] aimed at understanding the dimensions of cultural change, sharing learnings for challenges, and classifying repercussions on infrastructures. To do so, we discussed case studies to identify conditions of successful cultural change, the role of infrastructures as enablers of cultural change, and developments that promoted or hindered the use of infrastructures for cultural change. Different case studies from health sciences, art history, zoology, astrophysics, and psychology were discussed along a set of questions. Based on these cases we observed conditions for success: The technological possibilities were welcomed by local staff, grounded in already established practices, furthered by management, and continued by permanent academic staff.

2.1 A Scheme for Collecting Use Cases in Cultural Change

Based on these workshops we have developed a scheme for collecting and describing use cases of cultural change. This scheme has been tested during our second workshop, in which we focused on the tangible feedback of infrastructures and the actors, practices, and processes involved in cultural change. If you have a use case, we would be happy if you would share it with us. You can find printed forms at our poster at the conference, or you can follow this QR code to an online form at <https://cloud.nfdi4culture.de/apps/forms/s/Ej4gLA6zZ8boqtn5g9fTaGPa>.



Figure 2: QR code to our digital form for collecting use cases (<https://cloud.nfdi4culture.de/apps/forms/s/Ej4gLA6zZ8boqtn5g9fTaGPa>)

3. The Way Ahead

The initial workshops revealed that the differences between research communities are often less pronounced than the differences between individual scientists. Therefore, the collection of use cases of cultural change from communities provides a basis to develop learnings for all NFDI consortia and derive approaches to initiate cultural change processes. The learnings from our use cases will be discussed during a final workshop in Duisburg in November 2023 and published as a white paper at the end of the year.

Data availability statement

not applicable.

Underlying and related material

not applicable.

Author contributions

Julia Rakers, Bernhard Miller, Julia Mohrbacher, Daniel Nüst, Torsten Schrade, Jörg Seegert, and Christian Vater have contributed to **writing this contribution**, Holger Simon, and Cord Wiljes have contributed by **conceptualizing and/or editing**.

Competing interests

The authors declare no competing interests.

Funding

This contribution is funded by DFG through various grants:

KonsortSWD – grant no. 442494171

NFDI4Culture – grant no. 441958017

NFDI4Earth – grant no. 460036893

NFDI-MatWerk – grant no. 460247524

Acknowledgement

-

References

1. [FAIR] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
2. [BLV] Bund-Länder-Vereinbarung zu Aufbau und Förderung einer nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018. <https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/NFDI.pdf> (accessed on 2023-07-20)
3. [DSSS] Zenk-Möltgen, W., and Lepthien, G., "Data sharing in sociology journals", *Online Information Review*, Vol. 38 Issue: 6, pp.709-722, 2014, <https://doi.org/10.1108/OIR-05-2014-0119>
4. [TW] Van Tuyl, S., and Whitmire, A.L., "Water, Water, Everywhere: Defining and Assessing Data Sharing in Academia". *PLoS ONE* 11(2), 2016, <https://doi.org/10.1371/journal.pone.0147942>
5. [TUDelft] Teperek, M. and Dunning A., "Strategic Framework for Data Stewardship", 2019. <https://www.tudelft.nl/library/research-data-management/r/support/data-stewardship/support/strategic-framework-for-data-stewardship> (accessed on 2023-04-20)
6. [FDM] Thuringian Competence Network for Research Data Management, "Documents", 2023. https://forschungsdaten-thueringen.de/Info_Material_V02_en.html (accessed on 2023-04-21)
7. [CCBYUS1] Miller, B., Vater, C., Rakers, J., and Schrade, T., "CC-BY-US: Cultural Change beim Teilen von Forschungsdaten mit und durch die NFDI", 2023. <https://events.nfdi4culture.de/event/13/> (accessed on 2023-04-21)
8. [CCBYUS2] <https://events.nfdi4culture.de/event/25/> (accessed on 2023-07-20)
9. [CRedit] Allen, L., O'Connell, A., and Kiermer, V. (2019), "How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRedit) is helping the shift from authorship to contributorship". *Learned Publishing*, Vol. 32, pp. 71-74, 2019, <https://doi.org/10.1002/leap.1210>.
10. [N4E] Seegert, J., Nüst, D., and Bernard, L., "Initiating Cultural Change in the German Earth System Sciences Community with a Commitment Statement", *EGU General Assembly 2023*, Vienna, Austria, 24–28 Apr 2023, EGU23-14456, <https://doi.org/10.5194/egusphere-equ23-14456>.

Exploring and Improving Workflows for the Donation and Curation of Research Data

Natalie Kiesler¹[\[https://orcid.org/0000-0002-6843-2729\]](https://orcid.org/0000-0002-6843-2729), Daniel Schiffner¹[\[https://orcid.org/0000-0002-0794-0359\]](https://orcid.org/0000-0002-0794-0359)

¹DIPF Leibniz Institute for Research and Information in Education, Germany

Abstract: Managing research data and preparing it for long term archiving at a research data center is a time-consuming, repetitive and error-prone process without obvious rewards or incentives. To address these challenges, we aim for the integration of a standard data management tool into one of the research data centers of the German Network of Educational Research Data. The goal is to improve workflows for both, data donors and curators and to explore transferable solutions that allow researchers and curators to work on the same platform when entering and editing metadata.

Keywords: Research Data Management, Research Data Center Workflows, Efficiency, Interoperability, Controlled Vocabularies

1 Background and Motivation

Ever since research data management has become a core element of research projects across disciplines, the need for efficient workflows for donation and curation processes is on the increase. This is also true for the educational (technology) research communities [1], [2] and respective research data centers (RDCs) in Germany, such as those within the German Network of Educational Research Data (GNERD). A key objective of these RDCs is to develop and offer a research data infrastructure to process different types of data and metadata originating from educational research. However, data donation and curation processes usually require extensive effort to reach maturity (see, e.g., [3], [4]), resulting in multiple months of repetitive work for both, researchers and curators to enter, re-enter, verify, and double-check data and metadata. This high-threshold, unrewarded process is one of the well-known obstacles preventing researchers from donating data [5].

The motivation of the present work regards the delivery of research data management infrastructure within the GNERD. We aim at improving data donation and curation workflows of RDCs by exploring new tools, plugins, and interfaces for the exchange of data while ensuring interoperability. So the goal is to enhance the data donation process and the workload for the collection and curation of research data. This exploration and vision is transferable to other RDCs and disciplines.

2 Current Challenges among Donation and Curation Processes

Despite the availability of tools and other resources for the management and publication of research data (e.g., GitHub, OSF, Zenodo, DMPTool), the donation of research

data to one of the GNERD RDCs requires an extensive effort. This is mostly due to the fact that the RDCs require researchers to submit a rich set of metadata along with their actual research data. In addition, a technical report addressing the data collection and analysis is required. This is unfortunate due to several reasons. One of them is that researchers may have already collected all metadata within another data management tool that does not support the export in a format desired by certain RDCs. Hence, researchers have to repeat entering all metadata into another system with another metadata scheme, which results in an extensive, and error-prone process for both, researchers and curators. In addition, the effort remains unrecognized within the community [1], [2].

3 A Vision of New Workflows

Abandoning FAIR [6] and Open Data practices, however, is not an option. Therefore, we are exploring new ways to enable research data management for the community. As part of this process, we are exploring existing tools such as the Research Data Management Organiser (RDMO) with the goal of connecting it to one of GNERDs' RDCs. This way, researchers can manage their data during a project or study and have the data ready for export to an RDC as soon as the project ends, thereby supporting its full life cycle. Yet another advantage is the low threshold for researchers to donate data, thereby opening up options to provide metadata before concluding a research study.

In order to achieve these goals and long-term archiving, we have to integrate the RDC's internal data structure into an available data management tool. This is how both the management of research data for researchers and the curation of the data can be conducted via that same tool and user interface. Thus, metadata only needs to be entered once. At the same time, we want to use the same tool for the management of controlled vocabularies to simplify workflows.

Fortunately, tools like RDMO are available open source [7]. It offers support to researchers as it collects and structures metadata of research projects in the form of data management plans and so-called surveys. Moreover, a tool like RDMO offers machine actionable export options, and thus basic interoperability, e.g., for the exchange of data between researchers and an RDC.

As a first step towards this interoperability, we used existing code that generates machine actionable data management plans as a template [8]. We then adapted the structure and metadata scheme of the generated JSON file to represent those of the GNERD. Figure 1 contains an excerpt of such a newly generated JSON file with the GNERD's metadata scheme for projects. The next step was to map and align RDMO's standard data management plan with the metadata fields required by the GNERD.

One of our current challenges is related to the presentation of controlled vocabularies within RDMO so that curators at RDCs can also start working with RDMO. First of all, the list of available terms is extensive. So we decided to retrieve the basic controlled vocabularies and its terms from one of our APIs. Controlled vocabularies with multiple hierarchy levels are still work in progress, as their display in RDMO's GUI has to be determined. Another challenge is the mapping of all available vocabularies to the those of RDMO's standard data management plan.

```

1  {
2  "project": {
3    "title": "Metadata Export (work in progress)",
4    "subtitle": "In scheme of German Network of Educational Research Data",
5    "acronym": "XP-VFDB",
6    "duration_from": "2022-11-09",
7    "duration_to": "2023-05-31",
8    "funding_code": "DFG-0815",
9    "funding_agency": "DFG - Deutsche Forschungsgemeinschaft",
10   "programme": {
11     "acronym": "EBF"
12   },
13   "description": "Export a Data Management Plan from RDMO into GNERD's JSON scheme",
14   "comment_external": "(just a default comment)",
15   "involved_persons": [
16     {
17       "name": "Doe, John",
18       "orcid": "0000-0002-1234-5678",
19       "pnd_id": "https://explore.gnd.network/gnd/987654321",
20       "affiliation": {
21         "name": "Leibniz-Institut f\u00fcr Bildungsforschung und Bildungsinformation"
22       },
23       "person_function": {
24         "name": "Projektleitung"
25       }
26     }
27   ]
28 }
29

```

Figure 1. Excerpt of a new JSON file reflecting the GNERD metadata scheme for projects.

4 Conclusion

So far, adapting and extending RDMO's export options for the integration into the RDC seems promising. It proves RDMO's interoperability and its potential for a more efficient data management, donation, and curation in the educational (technology) research community. We will continue to pursue the goal of achieving a more efficient, and low-threshold data management, donation, and curation process where researchers and curators can operate on the same platform instead of multiple ones, while still providing the flexibility for each RDC to provide and gather the metadata they need. This new workflow can also avoid the error-prone re-entering of metadata within multiple tools or platforms. RDMO seems to be one of the tools that could be used for that purpose, and thus help to efficiently prepare research data and metadata for long-term archiving and secondary research.

Acknowledgements

We appreciate the contribution of our Software Engineer Axel Nieder-Vahrenholz to the exploration and integration of the Research Data Management Organizer.

Competing Interests

The authors declare they have no competing interests.

References

- [1] N. Kiesler and D. Schiffner, "On the lack of recognition of software artifacts and its infrastructure in educational technology research," in *20. Fachtagung Bildungstechnologien (DELFI)*, P. A. Henning, M. Striewe, and M. Wölfel, Eds., Bonn: Gesellschaft für Informatik e.V., 2022, pp. 201–206. [Online]. Available: <https://doi.org/10.18420/delfi2022-034>.

- [2] N. Kiesler and D. Schiffner, "Why We Need Open Data in Computer Science Education Research," in *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education Vol. 1*, ser. ITiCSE 2023, Turku, Finland: Association for Computing Machinery, 2023, ISBN: 979-8-4007-0138-2/23/07. DOI: [10.1145/3587102.3588860](https://doi.org/10.1145/3587102.3588860).
- [3] N. Kiesler, *Dataset: Recursive problem solving in the online learning environment CodingBat by computer science students*, Online, Datenerhebung: 2017. Version: 1.0.0. Datenpaketzugangsweg: Download-SUF. Hannover: FDZ-DZHW. Datenkuratierung: İköz-Akıncı, Dilek, Jun. 2022. DOI: <https://doi.org/10.21249/DZHW:studentsteps:1.0.0>.
- [4] N. Kiesler, "Recursive problem solving in the online learning environment CodingBat by computer science students," Tech. Rep., Jun. 2022. [Online]. Available: [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-studentsteps\\$/attachments/studentsteps_Data_Methods_Report_de.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-studentsteps/$/attachments/studentsteps_Data_Methods_Report_de.pdf).
- [5] C. L. Borgman and I. V. Pasquetto, *Why data sharing and reuse are hard to do*, 2017. [Online]. Available: <https://escholarship.org/uc/item/0jj17309>.
- [6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [7] J. Klar, *Research data management organiser*, 2023. [Online]. Available: <https://github.com/rdmorganiser/rdmo>.
- [8] T. Miksa, *Rda-dmp-common-standard*, 2020. [Online]. Available: <https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard>.

Zarr

A Cloud-Optimized Storage for Interactive Access of Large Arrays

Josh Moore¹[\[https://orcid.org/0000-0003-4028-811X\]](https://orcid.org/0000-0003-4028-811X), Susanne Kunis²[\[https://orcid.org/0000-0001-6523-7496\]](https://orcid.org/0000-0001-6523-7496)

¹ German BioImaging – Gesellschaft für Mikroskopie und Bildanalyse e.V.

² Department of Biology/Chemistry and Center for Cellular Nanoanalytics, University Osnabrück, Germany

Abstract. For decades, the sharing of large N-dimensional datasets has posed issues across multiple domains. Interactively accessing terabyte-scale data has previously required significant server resources to properly prepare cropped or down-sampled representations on the fly. Now, a cloud-native chunked format easing this burden has been adopted in the bioimaging domain for standardization. The format — Zarr — is potentially of interest for other consortia and sections of NFDI.

Keywords: FAIR, Community, Bioimaging, Data, Cloud, Format

In an ideal FAIR [1] bioimaging world, the seamless sharing of large image data — dense, often terabyte-scale, N-dimensional arrays — from microscope through to publication and even re-analysis would be possible (Figure 1). The reality, unfortunately, is much less clear due to the lack of a common format for exchange.

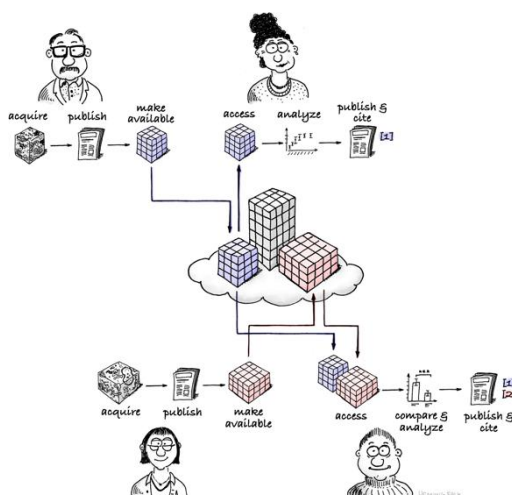


Figure 1. FAIR sharing of data is beneficial for both data producers and consumers. Consumers gain access to interesting datasets that would otherwise be out of reach. Producers get citations to their work when consumers publish their derivative work. OME-Zarr is the technology basis for enabling effective FAIR sharing of large image datasets. "FAIR re-use" by Henning Falk, ©2022 NumFOCUS, is used under a CC BY 4.0 license. [2]

Over 150 different file formats are produced by acquisition systems. Often the data must be duplicated into one or more additional formats for specific applications, increasing storage requirements. Libraries like Bio-Formats [3] can be used to extract the pixel data as well as critical metadata, making it possible to develop server systems like OMERO [4] and the Image Data Resource (IDR) [5]. However, users who would like to download the data are left with the often-complicated translation burden.

To provide a common container while enabling cloud-optimized sharing, the Open Microscopy Environment (OME) began the development of a next-generation file format (NGFF) in 2018 which was subsequently published in 2021 [6]. The basis of this work is Zarr (<https://zarr.dev>), a format specification for the storage of large N-dimensional typed arrays which avoids certain scalability issues by chunking data into atomic units which can be written and read independently.

Based closely on the HDF5 model [7], Zarr stores hierarchical groups of datasets with arbitrary metadata attached at each level of the structure. Critically, Zarr differs from HDF5 in that rather than using a complex internal binary data structure, a Zarr dataset comprises many individual files so that each chunk or metadata file (written in JSON) can be referenced via predefined, externally stable paths which can be listed by standard file and web browsers (Figure 2). For many of the storage backends, this enables the parallel writing of large image datasets, essential for cluster and cloud-based processing, as well as the viewing of terabytes of data from a static webpage. Implementations exist in several programming languages including C++, Java, JavaScript, Julia, Python, R, and Rust.

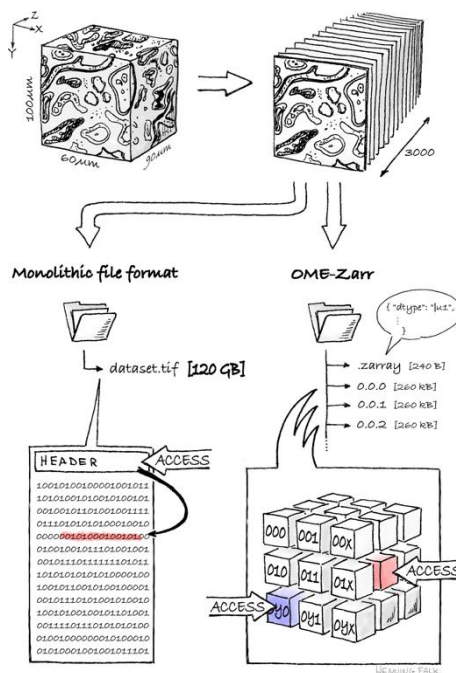


Figure 2. Technical differences between monolithic and chunked file formats. "Monolithic vs. chunked" by Henning Falk, ©2022 NumFOCUS, is used under a CC BY 4.0 license. [2]

By adding OME metadata [8] to the Zarr files, a standard imaging format has been created, supported by multiple visualization tools available with more in development (Figure 3). The flexible metadata structure permits the addition of further schemas like "Recommended Metadata for Biological Images" (REMBI) [9], "Quality Assessment and Reproducibility for Instruments & Images in Light Microscopy" (QUAREP-LiMi) [10], or "Minimum information guidelines for highly multiplexed tissue images" (MITI) [11] in the future. Repositories like the Bi-

olmage Archive [12] with the stated goal of accepting all published image data need the scalability afforded by OME-Zarr's serverless capabilities, while the format is also ideal for storing datasets like ZebraHub [13] or the Cell Painting Gallery [14], on institute storage or in Amazon's Open Data program respectively. More information on this growing ecosystem is available in a recent preprint on OME-Zarr [15].



Figure 3. Several visualization tools already support OME-Zarr, including Fiji/Bigdata-viewer/MoBIE [16] on the desktop, webKnossos [17] and Neuroglancer [18] on the web. With OME-Zarr, a selection of software tools and devices can access the same datasets from centralized storage. "Multiple clients" by Henning Falk, ©2022 NumFOCUS, is used under a CC BY 4.0 license. [2]

The technical issue which led to the development of Zarr, however, affects other communities. There is significant latency inherent in working with remote data, especially on cloud storage. In the original OME-NGFF paper [6], a benchmark compared the performance of Zarr, TIFF (Tagged Image File Format), and HDF5 files containing the same synthetic data accessed locally, via HTTP, and S3. The benchmark sequentially loaded random, uncompressed chunks to identify the overhead experienced by users visualizing data. It was shown that for monolithic formats like TIFF and HDF5 the overhead of remotely traversing the internal binary structure leads to performance degradation. The pre-computable locations of the Zarr chunks and metadata, in comparison, scale more favorably as latency increases. This is not to say that Zarr is fundamentally preferable to HDF5, but rather that there is a second, cloud-native paradigm with distinct characteristics (Table 1) which need to be considered when sharing large arrays and when developing software infrastructure.

Table 1. Back of the envelop comparison for characteristics of the two primary storage domains – filesystems and object storage, adapted originally from <https://www.openio.io/blog/block-file-object-storage-evolution-computer-storage-systems>

	Filesystem	Object storage ("cloud")
Storage cost	1 EUR/GB	0.01 EUR/GB
Throughput	Gbps	Tbps
Latency	10 μ s	1 ms
Modifications	I/O intensive	Immutable

To this end, GUIs and libraries are being updated or created, which multiple institutes are using to migrate their data to OME-Zarr and publish it online. In our recent preprint [15], members of the bioimaging community seek to signal a clear investment in this Zarr-based format, both to industry partners as well as nearby communities. Discussion and the exchange

of specifications is already occurring, e.g., with the geospatial community. The Open Geospatial Consortium (OGC) has adopted Zarr as a community standard [19] while NASA divisions like POWER (<https://power.larc.nasa.gov/>) have migrated their currently NetCDF-based data to Zarr and other divisions are in the process of evaluating the solution. We expect this process to continue as this becomes an agreed upon mechanism for sharing data moving forward.

Task Area 1, “Image (meta)data formats & standardization”, of the NFDI4BIOIMAGE consortium is committed to delivering a Zarr-based format for bioimaging, but as a general purpose, multi-language, extensible yet approachable specification, Zarr could be of interest to other NFDI consortia and the sections.

Competing interests

J.M. is a member of the steering councils of both the OME and Zarr software projects and holds equity in Glencoe Software, a commercial company that builds, delivers, supports, and integrates image data management systems across academic, biotech, and pharmaceutical industries.

Funding

J.M. was supported for work on OME-NGFF by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) NFDI4BIOIMAGE project 501864659 and by Chan Zuckerberg Initiative DAF grant numbers 2019-207272 and 2022-310144 as well as for work on Zarr by Chan Zuckerberg Initiative DAF grant numbers 2019-207338 and 2021-237467.

Acknowledgement

J.M. would like to thank the international NGFF community (<https://ngff.openmicroscopy.org>) for the specifications, tools, data resources, and ever lively discussions.

References

1. M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” *Sci Data*, vol. 3, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
2. J. Moore, zarr-developers/zarr-illustrations-falk-2022: Zarr illustrations by Henning Falk (August 2022). 2022. doi: 10.5281/zenodo.7037679.
3. M. Linkert et al., “Metadata matters: access to image data in the real world,” *J. Cell Biol.*, vol. 189, no. 5, pp. 777–782, May 2010, doi: 10.1083/jcb.201004104.
4. J.-M. Burel et al., “Publishing and sharing multi-dimensional image data with OMERO,” *Mamm. Genome*, vol. 26, no. 9–10, pp. 441–447, Oct. 2015, doi: 10.1007/s00335-015-9587-6.
5. E. Williams et al., “The Image Data Resource: A Bioimage Data Integration and Publication Platform,” *Nat. Methods*, vol. 14, no. 8, pp. 775–781, Aug. 2017, doi: 10.1038/nmeth.4326.
6. J. Moore et al., “OME-NGFF: a next-generation file format for expanding bioimaging data-access strategies,” *Nat. Methods*, vol. 18, no. 12, pp. 1496–1498, Dec. 2021, doi: 10.1038/s41592-021-01326-w.
7. The HDF Group, Hierarchical Data Format version 5. 1997-2021. [Online]. Available: <http://www.hdfgroup.org/HDF5>
8. I. G. Goldberg et al., “The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging,” *Genome Biol.*, vol. 6, no. 5, p. R47, May 2005, doi: 10.1186/gb-2005-6-5-r47.

9. U. Sarkans et al., "REMBI: Recommended Metadata for Biological Images-enabling reuse of microscopy data in biology," *Nat. Methods*, vol. 18, no. 12, pp. 1418–1422, Dec. 2021, doi: 10.1038/s41592-021-01166-8.
10. G. Nelson et al., "QUAREP-LiMi: A community-driven initiative to establish guidelines for quality assessment and reproducibility for instruments and images in light microscopy," *arXiv [q-bio.OT]*, Jan. 21, 2021. [Online]. Available: <http://arxiv.org/abs/2101.09153>
11. D. Schapiro et al., "MITI minimum information guidelines for highly multiplexed tissue images," *Nat. Methods*, vol. 19, no. 3, pp. 262–267, Mar. 2022, doi: 10.1038/s41592-022-01415-4.
12. M. Hartley, G. Kleywegt, A. Patwardhan, U. Sarkans, J. R. Swedlow, and A. Brazma, "The BioImage Archive - home of life-sciences microscopy data," *bioRxiv*, p. 2021.12.17.473169, Dec. 21, 2021. doi: 10.1101/2021.12.17.473169.
13. M. Lange et al., "Zebrahub – Multimodal Zebrafish Developmental Atlas Reveals the State Transition Dynamics of Late Vertebrate Pluripotent Axial Progenitors," *bioRxiv*, p. 2023.03.06.531398, Mar. 07, 2023. doi: 10.1101/2023.03.06.531398.
14. S. N. Chandrasekaran et al., "Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations," *bioRxiv*, p. 2022.01.05.475090, Jan. 05, 2022. doi: 10.1101/2022.01.05.475090.
15. J. Moore et al., "OME-Zarr: a cloud-optimized bioimaging file format with international community support," *bioRxiv*, Feb. 2023, doi: 10.1101/2023.02.17.528834.
16. C. Pape et al., "MoBIE: a Fiji plugin for sharing and exploration of multi-modal cloud-hosted big image data," *Nat. Methods*, vol. 20, no. 4, pp. 475–476, Apr. 2023, doi: 10.1038/s41592-023-01776-4.
17. K. M. Boergens et al., "webKnossos: efficient online 3D data annotation for connectomics," *Nat. Methods*, vol. 14, no. 7, pp. 691–694, Jul. 2017, doi: 10.1038/nmeth.4331.
18. J. Maitin-Shepard et al., *google/neuroglancer*: Zenodo, 2021. doi: 10.5281/ZENODO.5573293.
19. Open Geospatial Consortium. "Zarr Storage Specification 2.0 Community Standard". <https://portal.ogc.org/files/100727> (April 22, 2023)

NFDI4Earth: Improving Research Data Management in the Earth System Sciences

Lars Bernard^[<https://orcid.org/0000-0002-3085-7457>], Christin Henzen^[<https://orcid.org/0000-0002-5181-4368>], Auriol Degbelo^[<https://orcid.org/0000-0001-5087-8776>], Daniel Nüst^[<https://orcid.org/0000-0002-0024-5046>], and Jörg Seegert^[<https://orcid.org/0000-0001-9357-2830>]

Chair of Geoinformatics, TU Dresden, Germany

Keywords: Geospatial resources, Research Data Management, Earth System Sciences

The increasing availability of digital research products – typically encompassing data and software – in the Earth System Sciences (ESS) calls for new approaches to facilitate the management and reuse of these digital resources. NFDI4Earth addresses this gap and targets the harmonization of services related to digital research products in the ESS. In our initial phase, we aim to support researchers in (i) discovering and exploring relevant data and software sources, (ii) exploratory spatial data analysis, (iii) solving research data management problems and (iv) creating and publishing information products [1, 2]. This abstract briefly presents concepts and first results of the harmonization efforts within NFDI4Earth. It touches upon four points: the harmonization of resource descriptions, the harmonization of ESS-specific support for research data management (RDM), joint training activities for researchers in the ESS, and the harmonization of discovery and publishing workflows.

Harmonizing resource descriptions: A common thread of the various disciplines within the ESS is the collection, processing and sharing of spatial data spanning over broad range of topics (e.g., climate, geology, marine, land use and property, habitats, etc.). Spatial data refers data to which a spatial position can be directly/indirectly assigned through the use of spatial references. These spatial data come in different formats, and so do the interfaces of the services used for their storage, processing and analysis. This diversity demands strategies for harmonizing and linking to established ESS standards for metadata, data and services such as the APIs from the Open Geospatial Consortium (<https://www.ogc.org/>). NFDI4Earth proposes two key innovations in this context: 1) the integration of various metadata from and for the ESS through the use of Semantic Web concepts (e.g., the Resource Description Framework) and bridging towards concepts such as FAIR Digital Objects; and 2) the NFDI4Earth Label as a digital badge to inform about the level of FAIRness of ESS services. As of this writing, the NFDI4Earth Knowledge Hub is implemented as an open-source software stack using a data management system and a tightly coupled triplestore. (Meta) data is harvested from different external sources (e.g., re3data for ESS repositories, Research Organization Registry and Wikidata for organizations, Digital Curation Centre for ESS-specific metadata standards, GitLab for software projects...) and enriched with internally collected information (Figure 1). Also, community-driven tools developed within the NFDI4Earth Pilots and Incubators (e.g., for data cube visualization) are made accessible through the Knowledge Hub. We thereby enable the ESS community to answer previously unanswered and challenging questions such as ‘list all ESS services related to Oceanography published within the last two years, along with their service types’.

Harmonizing RDM support: Researchers and research data managers face several challenges across the research data management life cycle (see e.g., [3]). These include, for

instance, questions related to what data can be legally and ethically collected and stored, compliance with funding agencies' requirements, and dilemmas related to the sharing of sensitive research data. While efforts at specific universities and university libraries are underway to facilitate RDM in-house services for the institutes' researchers, there is still a lack of platforms in the ESS where researchers can ask questions and receive help from their peers directly. This gap is addressed in NFDI4Earth through the User Support Network, which is currently envisioned as a single point of access to expert knowledge and experts for RDM-related issues. As the experts are distributed across the NFDI4Earth community, the most suitable person is matched to the inquiry independent from organizational membership, personal contacts, or previous public record on a topic.

Joint training activities: Through two-year training programs for doctoral and post-doctoral researchers, the NFDI4Earth Academy offers a learning environment within the ESS that goes beyond institutional boundaries. As of this writing, the Academy offers training to 39 Fellows from 24 institutions in Germany. The training features, among others, research data management and Data Science skills, "think tank" events to brainstorm and develop synergies between the fellows' projects, and cross-consortia activities (e.g., Hackathons).

Harmonizing discovery and publication workflows: Scientific (geo)data infrastructures, despite being sometimes established for years and using open standardized APIs, still lack shared policies and a common organizational framework for the sharing of scientific geo-spatial resources [4]. In ESS, we can build upon various well-known and accepted services. However, with the increasing number of diverse services, e.g., for specific tasks or formats, researchers lack an overview of existing services and support in choosing proper services according to their needs. Here, NFDI4Earth proposes two solutions for harmonizing discovery and publishing workflows: 1) a single access point, called OneStop4All, for the discovery of distributed resources and 2) a concept to publish (meta) data using best-fitting existing repositories or platforms. The custom open-source OneStop4All solution requests resource descriptions from the Knowledge Hub via an API to generate a user-friendly view of harmonized resource descriptions with particular respect to spatio-temporal characteristics (Figure 1). With the OneStop4All, we also implement the NFDI4Earth concept of data publishing, that is, recommending existing repositories where researchers can publish (meta) data for their resources as opposed to offering a new publishing platform. By doing so, we harmonize publication workflows building on community-accepted and established services, foster the visibility/transparency of relevant data outside NFDI4Earth, reduce implementation efforts in the project and support consolidation and efficient use of resources. Moreover, we envision fostering a cultural change in ESS, e.g., by focusing on FAIR, open and sustainable ESS services, supported by a NFDI4Earth Commitment Statement [5].

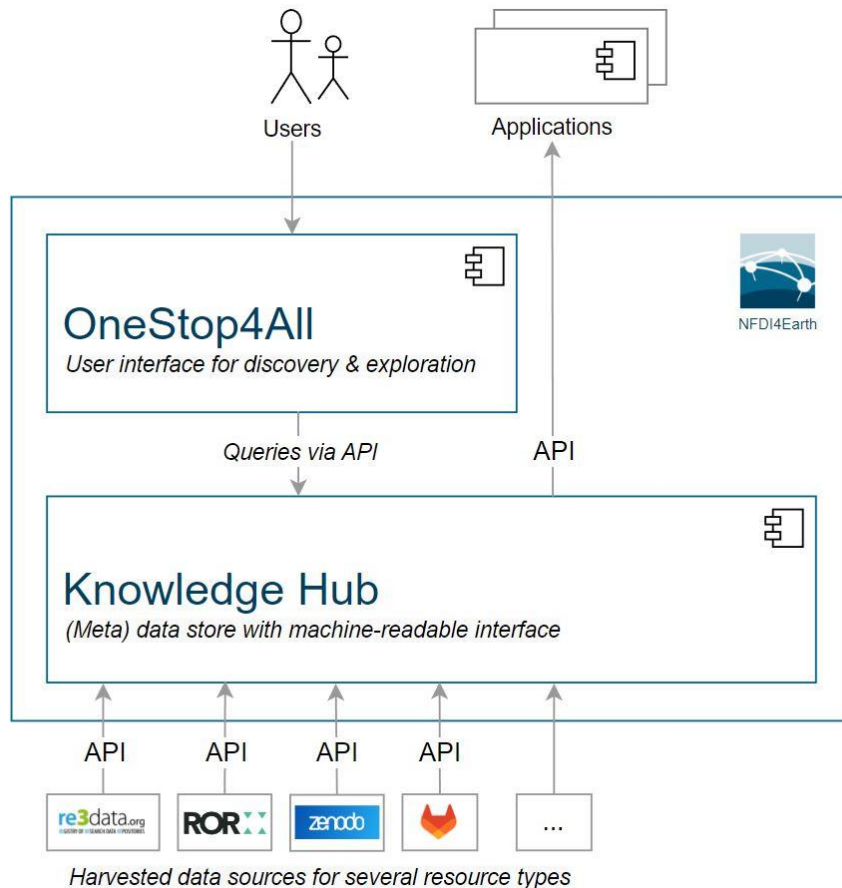


Figure 1. Simplified version of the NFDI4Earth Software Architecture

Data availability statement

The submission is not based on data.

Author contributions

L.B. contributed conceptualization, funding acquisition, supervision, review & editing. C.H. contributed conceptualization and writing – original draft. A.D. contributed conceptualization and writing – original draft. D.N. contributed conceptualization and writing – review & editing. J.S. contributed conceptualization, funding acquisition, project administration.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been funded by the German Research Foundation (DFG) through the project NFDI4Earth (DFG project no. 460036893, <https://www.nfdi4earth.de>) within the German National Research Data Infrastructure (NFDI, <https://nfdi.de/>).

Acknowledgement

The authors are grateful for the valuable input of the NFDI4Earth Consortium in the process of shaping and implementing these ideas.

References

1. L. Bernard, P. Bräsicke, R. Bertelmann, S. Frickenhaus, H. Gödde, C. Keßler, S. Lorenz, M. Mahecha, H. Marschall, D. Hezel, W. Nagel, M. Reichstein, M. Sester, H. Thiemann, C. Weiland, A. Wytzisk-Arens, NFDI4Earth Consortium, NFDI Consortium Earth System Sciences - Proposal 2020 revised, <https://doi.org/10.5281/zenodo.5718944>, 2021
2. C. Henzen, A. Degbelo, D. Nüst, Challenges in Developing a Software Architecture for a National Research Data Infrastructure in Earth System Sciences, EGU General Assembly Vienna, Austria, 24–28 Apr 2023, EGU23-17491, <https://doi.org/10.5194/egusphere-egu23-17491>, 2023
3. J. Bossaller, A.J. Million, The research data life cycle, legacy data, and dilemmas in research data management, Journal of the Association for Information Science and Technology, 1-6, <https://doi.org/10.1002/asi.24645>, 2022
4. L. Bernard, S. Mäs, M. Müller, C. Henzen, J. Brauner, Scientific geodata infrastructures: challenges, approaches and directions, International Journal of Digital Earth, 7(7), 613-633, <https://doi.org/10.1080/17538947.2013.781244>, 2014
5. J. Seegert, D. Nüst, L. Bernard, Initiating Cultural Change in the German Earth System Sciences Community with a Commitment Statement, EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, EGU23-14456, <https://doi.org/10.5194/egusphere-egu23-14456>, 2023.

Digital Twin-Based Concept for Reliable Research Data Management

Integrating Proprietary Data Sources for Hyperspectral Imaging

Alessa Rache^{1,2}[\[https://orcid.org/0000-0001-7598-8672\]](https://orcid.org/0000-0001-7598-8672),
Tim Häußermann^{1,2}[\[https://orcid.org/0000-0002-4020-4089\]](https://orcid.org/0000-0002-4020-4089),
Joel Lehmann^{1,2}[\[https://orcid.org/0000-0001-8261-8362\]](https://orcid.org/0000-0001-8261-8362), and
Julian Reichwald^{1,2}[\[https://orcid.org/0000-0002-4809-5710\]](https://orcid.org/0000-0002-4809-5710)

¹Center for Mass Spectrometry and Optical Spectroscopy

²Mannheim University of Applied Sciences

Abstract: In data-intensive research, reliable management of research data is a major challenge. In the field of Mass Spectrometry Imaging, vast amounts of data are being acquired from mostly proprietary data sources. Consequently, hindering seamless data integration into Research Data Management systems. Without a data repository, the continuous generation of scientific knowledge and innovative research based on existing information is limited. Moreover, to maintain the value of data to researchers throughout and beyond its lifecycle, FAIR principles for reliable data management approaches must be applied. To enable the required data transmission, the Digital Twin paradigm can be considered a reliable solution. The conceptual implementation of a heterogeneous mass spectrometer generating hyperspectral images leverages the Digital Twin to overcome common data management problems in data-intensive research.

Keywords: Research Data Management, Research 4.0, FAIR, Digital Twin, Container Digital Twin, Cyber-Physical System, Knowledge Graph, Ontology

1 Introduction

Reliable management of research data is becoming increasingly important, especially when dealing with data-intensive research [1]. As such, Mass Spectrometry Imaging (MSI) uses a combination of molecular mass analysis and spatial distribution to study the allocation of molecules present in the samples examined, visually mapped by hyperspectral images [2], [3]. This allows the generation of mass spectra for each single spot and consequently the acquisition of thousands of individual mass spectra per examination [4]. Accordingly, laboratories utilizing MSI generate vast amounts of data, causing an urgent need for extensive efforts regarding Research Data Management (RDM) [5], [6].

The primary objective of RDM is to pave the way for new scientific knowledge and enable innovative research based on existing information [7]. In response to emerging

efforts to reform research communication systems, Force 11 established the FAIR Principles in 2016. These principles are intended to serve as a guide seeking to improve the reusability of data, according to which data is expected to be Findable, Accessible, Interoperable, and Reusable to maintain its intrinsic value to researchers throughout the entire data lifecycle and beyond [8].

In reality, research data is commonly stored on local computers or on offline data repositories, which raises major concerns about the reproducibility of scientific research results [9]. The lack of standardization among the numerous software and hardware vendors contributes to the prevailing handling of data, complicating seamless integration. In practice, proprietary data formats and a high degree of heterogeneity across different devices are common [8]. This reality constrains the establishment of reliable RDM and the subsequent development of a knowledge base of relevance for the research community [10].

Research is increasingly adopting techniques raised by Industry 4.0 gearing itself up for Research 4.0 [11]. As an innovative technology for data transmission, the Digital Twin (DT) can be seen as a secure data source as it mirrors physical devices into the digital world through a bilateral communication stream [12], thus enabling the digital use and management of data [13].

2 Conceptual Architecture

In the preliminary work [14], [15] we proposed an RDM infrastructure offering extensive capabilities for storing and preserving research data in accordance with FAIR criteria utilizing DTs as well as agent-based DTs.

Figure 1 depicts an extension of the concept to overcome the limitation of proprietary measuring devices impeding automated data aggregation. The concept is divided into four segments: Physical Twin Space, Digital Twin Space, RDM Core Space and Smart Application Space.

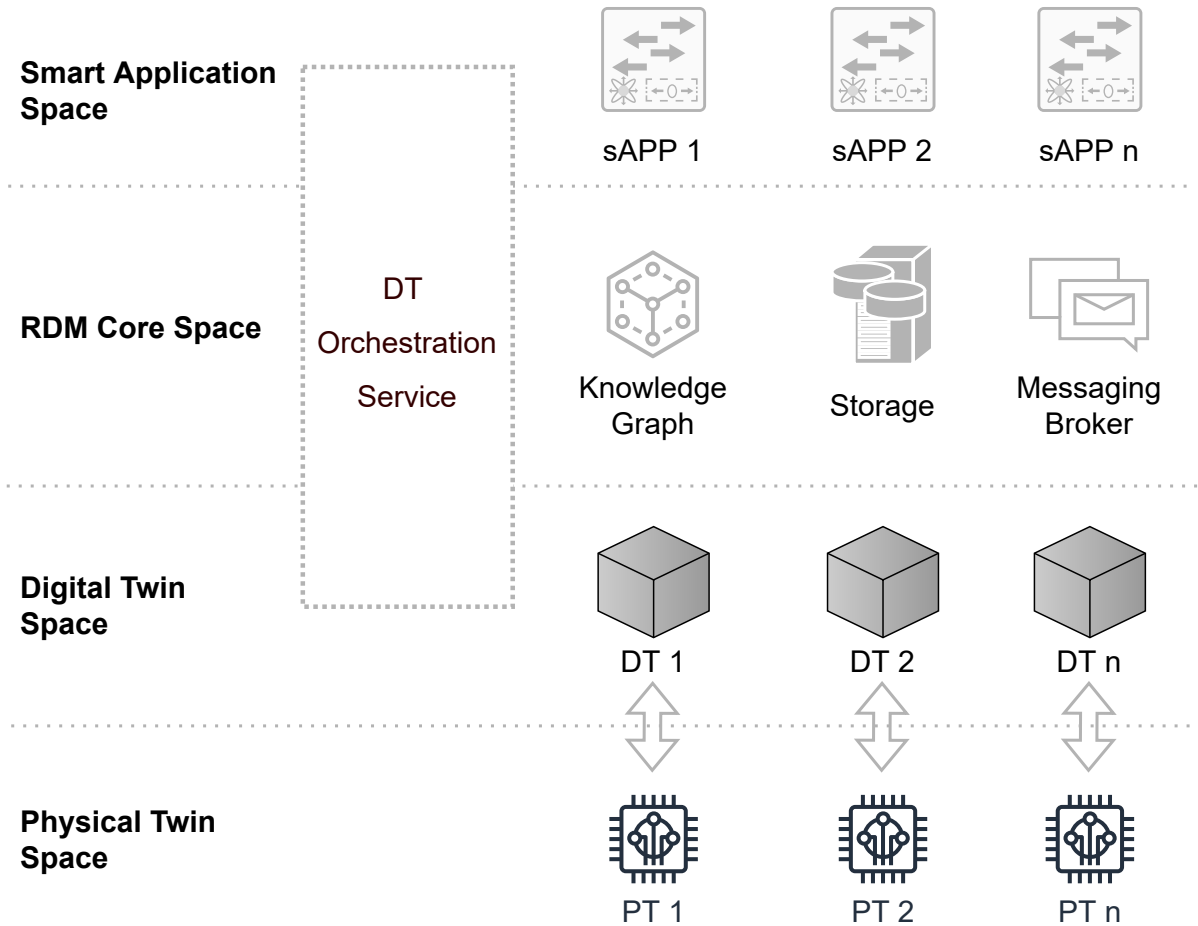


Figure 1. Conceptual RDM Architecture targeting proprietary devices without an open interface

The Physical Twin Space holds physical devices such as mass spectrometers ($PT_1 - PT_n$), each of having a unique DT ($DT_1 - DT_n$) which is provided using the existing infrastructure in the Digital Twin Space. Due to the proprietary nature of the devices, containerized DTs are utilized. These are distinguished by their portability, distributability and cloud independence from other types of DTs. As a result, they prevent interface- and accessibility-issues. For this purpose, the container-based DT is movable and thus can be run on the workstation which is associated to the corresponding physical device. Based on the lack of open interfaces, DTs cannot be instantiated by PTs themselves but must be obtained using a interface for specification. This interface is provided through the microservice DT Orchestration Service (DTOS). It provides the foundation for the execution of cross-layer tasks and communication. The DTOS deploys a container-based DT bound to the associated workstation, which then monitors the corresponding paths of relevant research data and handles the further storage procedures. The additions made to the DTOS and the use of the container-based DT constitute the extension of the architecture enabling connectivity and mapping of proprietary devices.

The workflow for the acquisition of hyperspectral data is as follows. Once the container-based DT is instantiated, the paths of the workstation are scanned continuously as new data is created. The container-based DT then recognizes the generation and launches a graphical user interface (GUI) on the workstation. The GUI includes several inputs needed to save the data in the Storage enriched by FAIR compliant metadata. After

completing all required entries, the container-based DT independently stores the data within the RDM architecture.

3 Conclusion

Conservatively decentralized handling of research data is attributed to prevailing proprietary data sources without open interfaces. Data-intensive research generates massive amounts of data which needs to be managed properly. To address this issue, a holonic infrastructure for reliable RDM using container-based DTs to integrate proprietary data sources is illustrated and exemplified by the research area of MSI.

The concept is generally applicable to any research area where seamless data transfer and reliable access to research data from proprietary physical data sources pose a major challenge such as environmental or physical sciences. The heterogeneous devices are mapped by container-based DTs that crawl and store research data into a central and structured FAIR-compliant RDM. Considering scientific knowledge generation and reusability, this concept opens up entirely new possibilities for exploiting the extensive potential of digitized RDM in research. The next logical step is to extend the concept to interdisciplinary research and the associated challenges, such as collaborative data use and related security concerns. To fully exploit the potential of the concept, a comprehensive implementation in the future is essential to evaluate its usability.

4 Appendix

Data availability statement

Not applicable.

Author contributions

Conceptualization, A.R., T.H., J.L.; methodology, A.R.; investigation, A.R., T.H., J.L.; writing-original draft preparation, A.R., T.H., J.L.; supervision, J.R.; project administration, A.R.; All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

Parts of this work presented in this paper were supported by a grant from the German Ministry of Education and Research (BMBF), grant number 16FDFH125.

References

- [1] J. Gray, D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber, "Scientific data management in the coming decade," *ACM SIGMOD Record*, vol. 34, no. 4, pp. 34–41, Dec. 2005, ISSN: 0163-5808. DOI: [10.1145/1107499.1107503](https://doi.org/10.1145/1107499.1107503). [Online]. Available: <https://doi.org/10.1145/1107499.1107503>.

- [2] E. R. Amstalden van Hove, D. F. Smith, and R. M. Heeren, "A concise review of mass spectrometry imaging," en, *Journal of Chromatography A*, vol. 1217, no. 25, Jun. 2010, ISSN: 00219673. DOI: [10.1016/j.chroma.2010.01.033](https://doi.org/10.1016/j.chroma.2010.01.033). [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0021967310000701> (visited on 12/13/2022).
- [3] F. Grélard, D. Legland, M. Fanuel, B. Arnaud, L. Foucat, and H. Rogniaux, "Esmraldi: Efficient methods for the fusion of mass spectrometry and magnetic resonance images," *BMC Bioinformatics*, vol. 22, no. 1, p. 56, Feb. 2021, ISSN: 1471-2105. DOI: [10.1186/s12859-020-03954-z](https://doi.org/10.1186/s12859-020-03954-z). [Online]. Available: <https://doi.org/10.1186/s12859-020-03954-z>.
- [4] A.-M. Lahesmaa-Korpinen, S. M. Carlson, F. M. White, and S. Hautaniemi, "Integrated data management and validation platform for phosphorylated tandem mass spectrometry data," *PROTEOMICS*, vol. 10, no. 19, 2010, ISSN: 1615-9861. DOI: [10.1002/pmic.200900727](https://doi.org/10.1002/pmic.200900727). [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.200900727>.
- [5] P. Romano, A. Profumo, M. Rocco, R. Mangerini, F. Ferri, and A. Facchiano, "Geena 2, improved automated analysis of MALDI/TOF mass spectra," *BMC Bioinformatics*, vol. 17, no. 4, p. 61, Mar. 2016, ISSN: 1471-2105. DOI: [10.1186/s12859-016-0911-2](https://doi.org/10.1186/s12859-016-0911-2). [Online]. Available: <https://doi.org/10.1186/s12859-016-0911-2>.
- [6] O. J. R. Gustafsson, L. J. Winderbaum, M. R. Condina, *et al.*, "Balancing sufficiency and impact in reporting standards for mass spectrometry imaging experiments," *GigaScience*, vol. 7, no. 10, Oct. 2018, ISSN: 2047-217X. DOI: [10.1093/gigascience/giy102](https://doi.org/10.1093/gigascience/giy102). [Online]. Available: <https://doi.org/10.1093/gigascience/giy102>.
- [7] A. Whyte and J. Tedds, "Making the Case for Research Data Management," *Digital Curation Centre Jisc Briefing Paper*, Sep. 2011. [Online]. Available: https://www.researchgate.net/publication/252931138_Making_the_Case_for_Research_Data_Management.
- [8] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," en, *Scientific Data*, vol. 3, no. 1, p. 160018, Mar. 2016, Number: 1 Publisher: Nature Publishing Group, ISSN: 2052-4463. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18). [Online]. Available: <https://www.nature.com/articles/sdata201618>.
- [9] M. Diepenbroek, F. O. Glockner, P. Grobe, *et al.*, "Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio)," p. 11, [Online]. Available: https://www.researchgate.net/publication/267574356_Towards_an_Integrated_Biodiversity_and_Ecological_Research_Data_Management_and_Archiving_Platform_The_German_Federation_for_the_Curation_of_Biological_Data_GFBio.
- [10] B. Mons, *Data Stewardship for Open Science: Implementing FAIR Principles*. New York: Chapman and Hall/CRC, Feb. 2018, ISBN: 978-1-315-38071-1. DOI: [10.1201/9781315380711](https://doi.org/10.1201/9781315380711).
- [11] E. Jones, N. Kalantery, and B. Glover, "Research 4.0: Interim report," en, Demos, Report, Oct. 2019. [Online]. Available: <https://apo.org.au/node/262636>.
- [12] M. Grieves, *Origins of the Digital Twin Concept*. Aug. 2016. DOI: [10.13140/RG.2.2.26367.61609](https://doi.org/10.13140/RG.2.2.26367.61609). [Online]. Available: https://www.researchgate.net/publication/307509727_Origins_of_the_Digital_Twin_Concept.
- [13] T. P. Raptis, A. Passarella, and M. Conti, "Data Management in Industry 4.0: State of the Art and Open Challenges," *IEEE Access*, vol. 7, pp. 97 052–97 093, 2019, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: [10.1109/ACCESS.2019.2929296](https://doi.org/10.1109/ACCESS.2019.2929296). [Online]. Available: <https://ieeexplore.ieee.org/document/8764545>.

- [14] J. Lehmann, S. Schorz, A. Rache, T. Häußermann, M. Rädle, and J. Reichwald, "Establishing reliable research data management by integrating measurement devices utilizing intelligent digital twins," *Sensors*, vol. 23, no. 1, 2023, ISSN: 1424-8220. DOI: [10.3390/s23010468](https://doi.org/10.3390/s23010468). [Online]. Available: <https://www.mdpi.com/1424-8220/23/1/468>.
- [15] J. Lehmann, A. Lober, T. Häußermann, et al., "The anatomy of the internet of digital twins: A symbiosis of agent and digital twin paradigms enhancing resilience (not only) in manufacturing environments," *Machines*, vol. 11, no. 5, 2023, ISSN: 2075-1702. DOI: [10.3390/machines11050504](https://doi.org/10.3390/machines11050504). [Online]. Available: <https://www.mdpi.com/2075-1702/11/5/504>.

Ten Simple Rules for Designing and Building a FAIR Research Infrastructure

Sharif Islam¹[\[https://orcid.org/0000-0001-8050-0299\]](https://orcid.org/0000-0001-8050-0299)

¹ Naturalis Biodiversity Center, Leiden, The Netherlands

Abstract. One of the key priorities of The European Strategy Forum on Research Infrastructures (ESFRI) is to build sustainable and FAIR (Findable, Accessible, Interoperable, Reusable) infrastructures. However, designing and building such infrastructures requires careful consideration of various factors, such as data interoperability, operational sustainability, and governance. This poster proposes ten simple rules, inspired by [Ten Simple Rules for scientific research](#), for designing and building a research infrastructure drawing from existing initiatives particularly from experiences in preparation of [DiSSCo](#) (Distributed System of Scientific Collections) – a new research infrastructure that was in the [ESFRI 2018 roadmap](#). While these rules are not comprehensive, they highlight a few essential traits that can be applied across different disciplines. For each rule, we highlight how within DiSSCo we accomplished the specific aspect.

Keywords: FAIR Digital Objects, Natural Science Collections

1. Introduction

One of the key priorities of The European Strategy Forum on Research Infrastructures (ESFRI) is to build sustainable and FAIR (Findable, Accessible, Interoperable, Reusable) infrastructures. However, designing and building such infrastructures requires careful consideration of various factors, such as data interoperability, operational sustainability, and governance. This poster proposes ten simple rules, inspired by [Ten Simple Rules for scientific research](#), for designing and building a research infrastructure drawing from existing initiatives particularly from experiences in preparation of [DiSSCo](#) (Distributed System of Scientific Collections) – a new research infrastructure that was in the [ESFRI 2018 roadmap](#). While these rules are not comprehensive, they highlight a few essential traits that can be applied across different disciplines. For each rule, we highlight how within DiSSCo we accomplished the specific aspect.

1.1 Rule 1: Ensure a Clear Purpose and Vision

Before designing and building a research infrastructure, it is important to have a clear purpose and vision. This should include a definition of the research infrastructure's scope, its target user community, and the expected outcomes. DiSSCo's focus on European natural science collections at the centre of data-intensive scientific excellence and innovation provided a clear purpose and vision to work towards [1].

1.2 Rule 2: Adopt FAIR Principles from the beginning

FAIR principles should be adopted from the beginning of the design process. This will enable the research community to find and use the data effectively, leading to discoveries and insights.

During the [ICEDIG](#) project (which delivered the DiSSCo design blueprint) and [DiSSCo Prepare](#), the FAIR implementation plan was at the forefront. DiSSCo's involvement in the ENVRI FAIR project helped in defining a FAIR implementation plan. DiSSCo's decision to use Digital Object Architecture also aligns with the ongoing work within the FAIR Digital Object specification [2, 3]. Looking into specific workflows from multiple organisations and understanding the data lifecycle (see Figure 1) helped us to scope different aspects of FAIR.

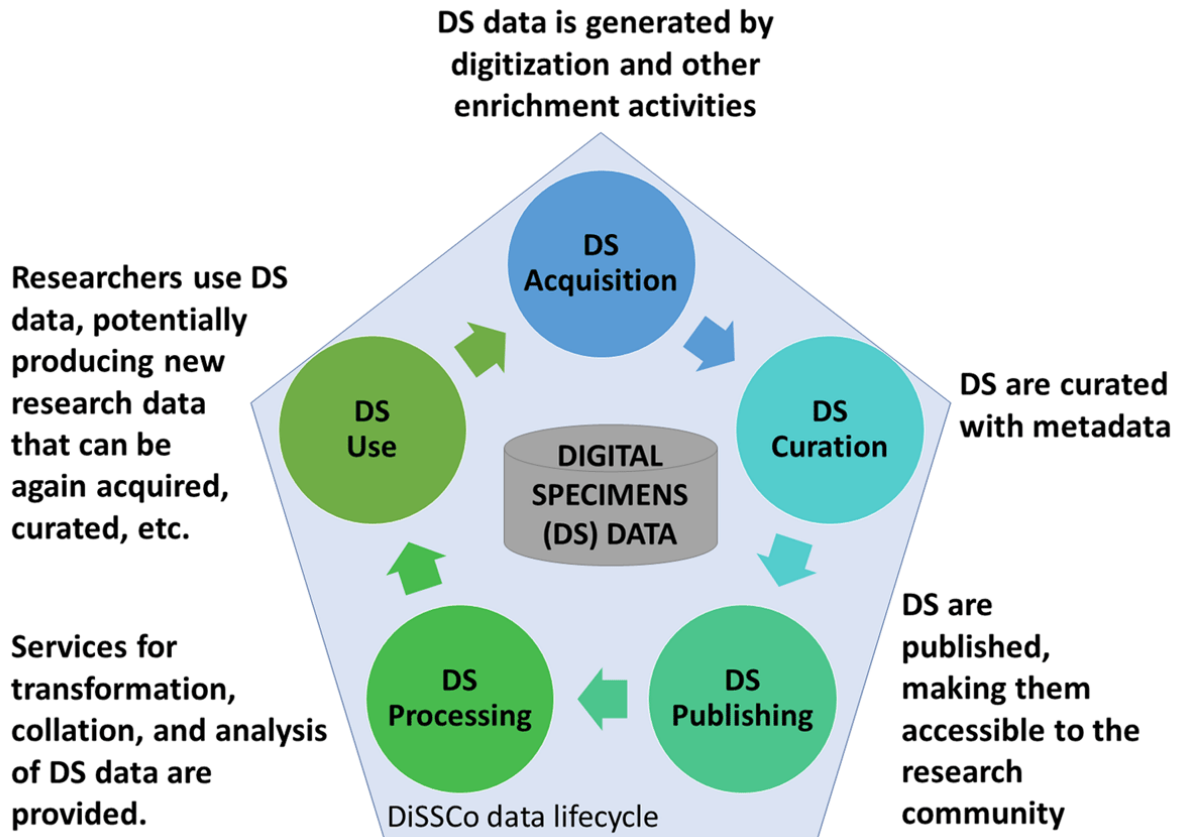


Figure 1: Digital Specimen (DS) lifecycle as the core data element for DiSSCo

1.3 Rule 3: Build on Existing Collaborations and Projects

By leveraging existing resources and expertise, it is possible to avoid duplicating efforts and to build on what has already been achieved. In DiSSCo's case, different digitisation projects over the past decades provided the foundation for building a data-driven ecosystem. EU funded Projects deliverables from ICEDIG, SYNTHEYES+ were valuable for the DiSSCo Prepare project. DiSSCo's involvement in other EU projects also provided opportunities to work and collaborate with other research infrastructures ([GBIF](#), [LifeWatch](#), [eLTER](#), [ELIXIR](#)).

1.4 Rule 4: Design for Interoperability

Interoperability is key to the success of any system. The infrastructure should be designed to enable the seamless exchange of data and metadata between different systems and platforms. This will ensure that the data is reusable and can be combined with other data to create new insights. DiSSCo's decision to use the FAIR Digital Object framework provides the building blocks for creating interoperable systems (see Figure 2; also see [4]) The current development effort also focuses on open source solutions that are better suited for delivering interoperable research infrastructure.

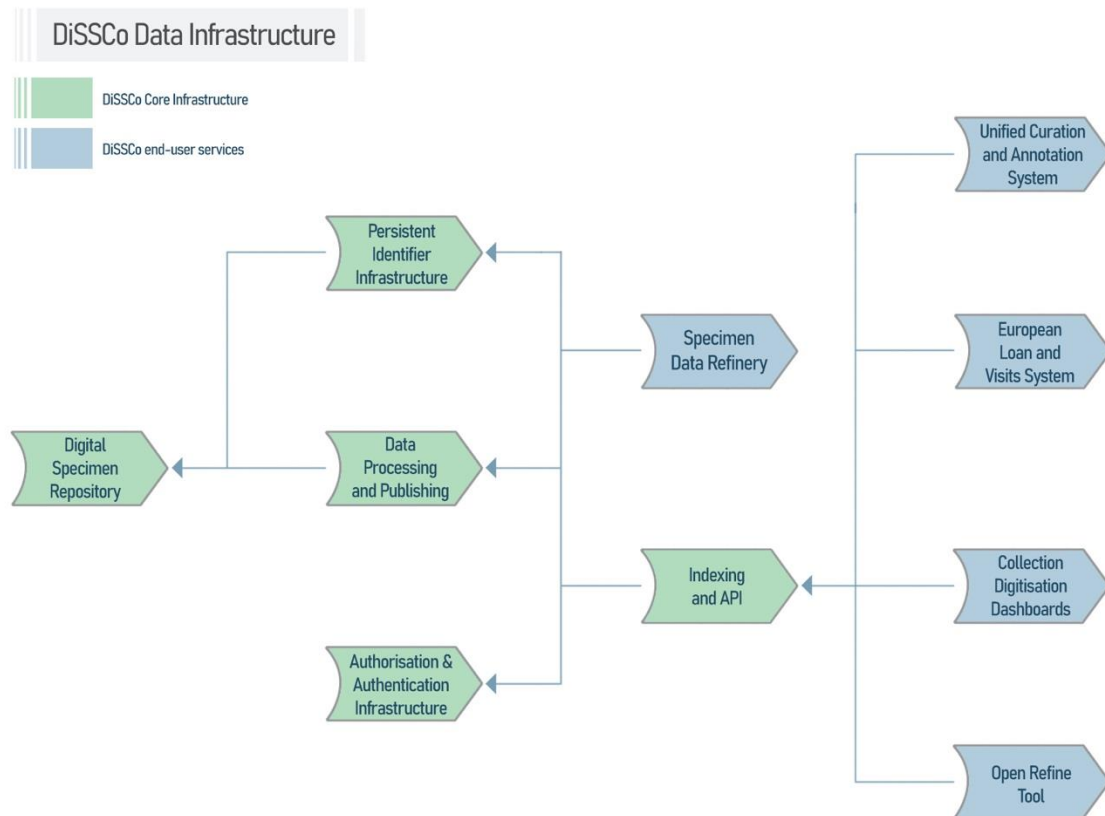


Figure 2: DiSSCo data infrastructure building blocks

1.5 Rule 5: Consider Sustainability

The infrastructure must be planned to guarantee long-term technical and financial viability. The technical team at DiSSCo adopted a framework for creating modular products that could be changed and expanded because of the use of FAIR principles, open source software, and the adoption of Agile and DevOps practices. Interoperability and sustainability also enable the system to be designed to be flexible and future-proof. For the financial aspect, extensive discussion and planning went on in the preparation phase of DiSSCo towards building a funding structure supported by the nation states and a plan towards forming an ERIC (European Research Infrastructure Consortium) [5].

1.6 Rule 6: Develop Data Governance and Policy Frameworks

The infrastructure should be designed to include policies and guidelines that govern data access, sharing, and reuse. This will ensure that the data is used ethically and in compliance with legal and ethical standards. Even though DiSSCo is not operational yet, our pilot is already considering the policy implications of different data workflows (from sample collections to publishing data). Some of these are also aligned with global policy discussions. At the same time, the aforementioned ERIC roadmap is contributing towards a robust discussion around governance mechanisms. These discussions have significant implications for how resources can be allocated to implement the wider technical vision.

1.7 Rule 7: Promote Openness and Collaboration

The infrastructure should be designed to promote openness and collaboration between different stakeholders, including researchers, data providers, and infrastructure providers. This will ensure that the infrastructure is used to its full potential and that discoveries and insights are created. During DiSSCo's design and preparation phase, we used various modes for collaboration and communication ([GitHub](#), newsletters, webinars for instance).

1.8 Rule 8: Plan for User Support and Training

The infrastructure should be designed to include user support services that help researchers to use the infrastructure effectively. This includes training, documentation, and technical support. Both SYNTHESYS+ and DiSSCo Prepare projects had a specific focus on user support and training.

1.9 Rule 9: Ensure Data Security and Privacy

The infrastructure should be designed to include security and privacy measures that protect sensitive data and comply with legal and ethical standards. Even though DiSSCo does not deal with human research data, we will have the researcher's profile and other personal information that needs to adhere to GDPR. Along with that, global procedures and steps that are in place need to be incorporated within DiSSCo for handling endangered species data.

1.10 Rule 10: Monitor and Evaluate Progress

Monitoring and evaluating progress can be achieved by regularly reviewing metrics such as usage and impact. The EU project framework for milestones and deliverables provides a structure for such monitoring and evaluation. However, DiSSCo internally also developed processes to ensure we are working towards our vision and allow opportunities for modification and improvement.

2. Conclusion

It is essential to recognise that the proposed set of ten "simple rules" for designing a FAIR research infrastructure, while inspired by the FAIR principles, extends beyond them. Rather than providing easy-to-follow rules for addressing the 16 FAIR principles, the poster focuses on a broader approach towards designing and building a sustainable and interoperable research infrastructure like DiSSCo. The findings emphasise that FAIR is not solely about making data discoverable, accessible, interoperable, and reusable. Instead, it encompasses a more holistic perspective, considering the entire data lifecycle, governance structure, and other digital objects (such as workflow management, algorithms, models, and research software [6]). This approach advocates for creating a modular, loosely coupled, yet integrated design that can serve the user community's needs.

The poster advocates for going beyond mere checkbox compliance with the FAIR principles and instead adopting a comprehensive, integrated approach. This forward-thinking, FAIR-by-design perspective will advance research, collaboration, and knowledge dissemination in various disciplines.

Author contributions

Sharif Islam [CRediT roles: Conceptualization, investigation, writing—original draft, review & editing].

Competing interests

The author declare that there is no competing interests.

Funding

The DiSSCo Prepare project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871043.

Acknowledgement

The authors thank and acknowledge valuable feedback offered by Wouter Addink, Eva Alonso, and the reviews.

References

1. A. Hardisty, H. Saarenmaa, A. Casino, M. Dillen, K. Gödderz, Q. Groom, H. Hardy, D. Koureas, A. Nieva de la Hidalga, D.L. Paul, V. Runnel, X. Vermeersch, M. van Walsum, and L. Willemse, "Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1," *Research Ideas and Outcomes*, vol. 6, p. e54280, 2020. [Online]. Available: <https://doi.org/10.3897/rio.6.e54280>
2. A.R. Hardisty, E.R. Ellwood, G. Nelson, B. Zimkus, J. Buschbom, W. Addink, R.K. Rabeler, J. Bates, A. Bentley, J.A. Fortes, and S. Hansen, "Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet," *BioScience*, vol. 72, no. 10, pp. 978-987, 2022. [Online]. DOI: <https://doi.org/10.1093/biosci/biac060>
3. L. Lannom, D. Koureas, and A.R. Hardisty, "FAIR data and services in biodiversity science and geoscience," *Data Intelligence*, vol. 2, no. 1-2, pp. 122-130, 2020. [Online]. DOI: https://doi.org/10.1162/dint_a_00034
4. T. Loo, M. Benvenuti, L. Cecchi, G. Innocenti, M. Biaggini, L. Bellucci, V. Moggi Cecchi, and F. Di Vincenzo, "DiSSCo Prepare Deliverable D9.6 - Compilation of Construction Masterplan," 2023. [Online]. DOI: <https://doi.org/10.34960/cy1m-b238>
5. S. Scory, C. Paleco, A. Casino, E. Alonso, D. Koureas, T. Loo, P. Mergen, A. Nivart, F. Dusoulier, and V. Demanoff, "MS-7.2 Analysis of the legal entity models and their suitability for achieving DiSSCo objectives," 2022. [Online]. URI: <https://know.dissco.eu/handle/item/473>
6. M. Barker, N.P. Chue Hong, D.S. Katz, et al., "Introducing the FAIR Principles for research software," *Sci Data*, vol. 9, p. 622, 2022. [Online]. DOI: <https://doi.org/10.1038/s41597-022-01710-x>

Ontology-Based Laboratory Data Acquisition with EnzymeML for Process Simulation of Biocatalytic Reactors

Alexander S. Behr¹[\[https://orcid.org/0000-0003-4620-8248\]](https://orcid.org/0000-0003-4620-8248), Elnaz Abbaspour¹[\[https://orcid.org/0009-0002-3335-6428\]](https://orcid.org/0009-0002-3335-6428),
Katrin Rosenthal²[\[https://orcid.org/0000-0002-6176-6224\]](https://orcid.org/0000-0002-6176-6224), Jürgen Pleiss³[\[https://orcid.org/0000-0003-1045-8202\]](https://orcid.org/0000-0003-1045-8202), and
Norbert Kockmann¹[\[https://orcid.org/0000-0002-8852-3812\]](https://orcid.org/0000-0002-8852-3812)

¹ Dept. of Biochemical and Chemical Engineering, TU Dortmund University, Germany

² School of Science, Constructor University, Bremen, Germany

³ Institute of Biochemistry and Technical Biochemistry, University of Stuttgart, Germany

Abstract. The presented work explores the use of ontologies and standardized enzymatic data to set up enzymatic reactions in process simulators, such as DWSIM. Setting up an automated workflow to start a process simulation based on enzymatic data obtained from the laboratory can help save costs and time during the development phase. Standardized conditions are crucial for accurate comparison and analysis of enzymatic data, where ontologies provide a standardized vocabulary and semantic relations between relevant concepts. To ensure standardized data, an electronic lab notebook (ELN) is used based on EnzymeML, an open standard XML-based format for enzyme kinetics data. Furthermore, two ontologies are merged and the result is extended for the use in the Python-based workflow. The resulting data is stored in a knowledge graph for research data in a machine-accessible and human-readable format. Thus, the study demonstrates a workflow that allows for the direct translation of ELN data into a process simulation via ontologies.

Keywords: Electronic Lab Notebooks, Enzymatic Catalysis, Knowledge Graph, Process Simulation

1. Introduction

The industrial production of biocatalytic processes has significant potential [1]. However, the development of new bioprocesses is a challenging and highly specific task about reaction conditions. Therefore, there is a high demand for tools, such as process simulators, that can help save costs and time during the process development phase [2]. For instance, the open-source process simulator DWSIM [3] enables the calculation of process streams prior to the establishment of the process in a laboratory plant.

Standardized conditions are crucial for comparing and operating on enzymatic data in bioreactors set up with enzymatic reactions in process simulators. Important parameters include enzyme activities and reaction kinetics, which can vary depending on specific reaction conditions. Furthermore, laboratory technicians who record enzyme-specific data may not be the same individuals who execute process simulations. To address these challenges, ontologies can be used as they provide a standardized vocabulary and semantic relations between concepts relevant to the research domains, enabling accurate comparison and analysis of enzymatic data [4, 5].

Furthermore, electronic laboratory notebooks (ELNs) help laboratory experimenters to record laboratory data and generating a machine-readable data collection while mitigating data loss. Thus, they enable clean research data management in laboratories. As ELNs exist in multiple shapes and utilize different formats, the focus in this work lies on the use of data stored in EnzymeML-based ELNs in form of pre-structured Excel-sheets [6, 7]. EnzymeML is not only an open standard XML-based format for enzyme kinetics data, but also uses ontology classes, e.g., from the Systems Biology Ontology (SBO) [8].

This work shows an automated approach to translate data contained in ELNs into a process simulation by standardized concepts stated in ontologies. Reading EnzymeML-based Excel-sheets with Python, data is extracted and stored in an ontology-based knowledge graph. Furthermore, data regarding the process flow sheet and additional data needed to setup the process simulation are included. Then, DWSIM and its Python interface are used to import the needed data for the automated setup of a process simulation of a reactor scale-up. After the process simulation is conducted, resulting data is also stored in the knowledge graph, allowing for automated storage of research data in a machine and human readable way. This overall workflow, allowing for direct translation of ELN data into a process simulation via ontologies, is depicted schematically in Figure 1.

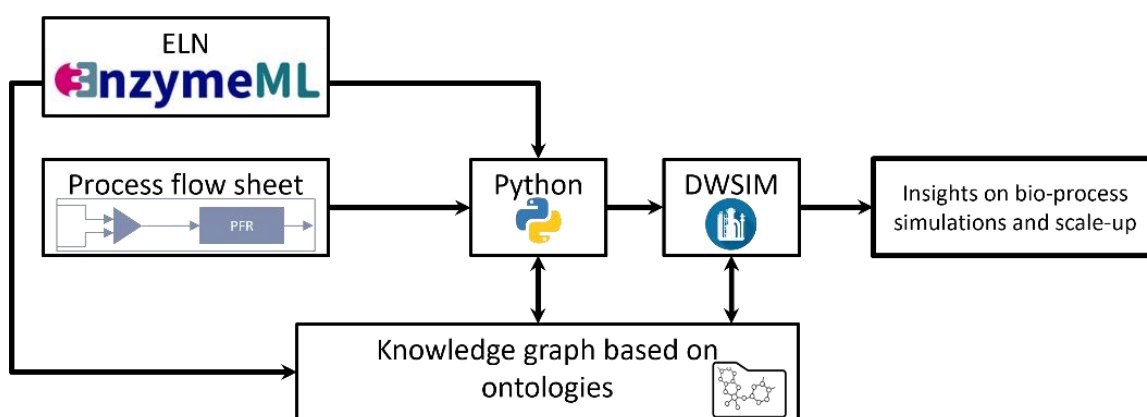


Figure 1. Schematic overview of the workflow presented in this work for the automated execution of process simulations based on a knowledge graph and ELN-based data.

2. Methods

To test the workflow, a set of ELN files describing experiments of oxidation of ABTS with the enzyme Laccase is used. The data of the experiments conducted in the lab in a continuous milliliter-reactor contain among others information of the reaction kinetics. Once, the EnzymeML-based ELN is filled in with the information of the laboratory experiments, the Python-package PyEnzyme [8] allows for automated data extraction from the ELN files.

In order to setup a knowledge graph for the experiments and the simulations in DWSIM, an ontology is needed. As the ELN is setup with classes from the SBO, it is used as base ontology. In addition, the metadata4ing ontology [9] is used to describe process-related concepts. Thus, classes from metadata4ing are included into the SBO and own classes and relations added where necessary to obtain an extended ontology tailored to the needs of this workflow.

Utilizing the owlready2 [10] module in Python, the ontology can then be loaded and extended automatically with the data from the ELN files. Figure 2 depicts an excerpt of the resulting knowledge graph, where individuals are assigned automatically to the ontology based on the input ELN data. This allows for structured access on, e.g., reaction kinetic parameters.

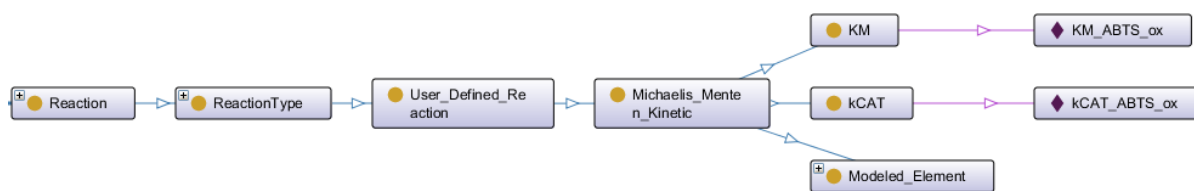


Figure 2. Excerpt of the ontology visualized in Protégé describing the class hierarchy (yellow circles) leading to the individuals for the kinetic parameters K_M and k_{Cat} of the Michaelis-Menten kinetics with regards to the substance $ABTS_{ox}$

Furthermore, the process simulator DWSIM is equipped with a Python-API allowing for automated setup of flow-sheets and execution of process simulations. Thus, after the knowledge graph is generated and stored, the information contained is transferred via the API, creating and executing a new process simulation.

3. Results

Executing the workflow described in the previous chapters results in a knowledge graph containing not only the ontology classes of the extended ontology, but also data obtained from the ELN. Figure 3 shows an excerpt of the class Laccase visualized in Protégé with the data annotations used to setup the corresponding substance in DWSIM. Furthermore, the resulting process simulation is visualized. Thus, this workflow allows for a quick and automated setup of process simulations based on laboratory data previously recorded in an EnzymeML-based ELN.

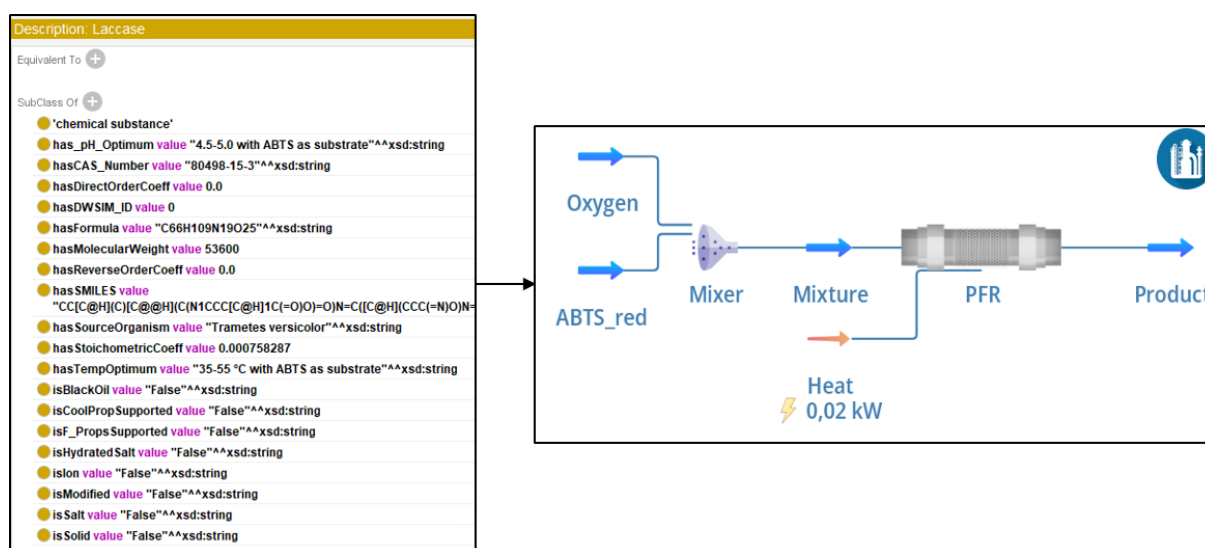


Figure 3. Excerpt of the annotation of the Enzyme Laccase within the knowledge graph visualized with Protégé (left) and resulting process simulation of a plug flow reactor (PFR) in DWSIM (right).

Data availability statement

The data, code and markdown files presented in this abstract will be available at GitHub here: <https://github.com/TUDoAD/EnzymeMLandDWSIM>

Author contributions

Conceptualization: A.S.B.; Methodology: A.S.B., E.A.; Software: E.A., A.S.B.; Validation: E.A., J.P., K.R.; Data Curation: K.R.; Writing – Original Draft: A.S.B., Writing – Review & Editing: N.K., J.P., K.R.; Visualization: A.S.B., E.A.; Supervision: A.S.B., N.K.

Competing interests

The authors declare that they have no competing interests.

Funding

The Deutsche Forschungsgemeinschaft (DFG) is acknowledged for funding this research as part of the Nationale Forschungsdateninfrastruktur (NFDI) initiative (grant No.: NFDI/2-1 - 2021).

Acknowledgement

The authors thank Julia Suhrkamp for the supervision of the laboratory experiments enabling the data recording with EnzymeML used in this work.

A.S.B. thanks the networking program 'Sustainable Chemical Synthesis 2.0' (SusChemSys 2.0) for the support and fruitful discussions across disciplines.

References

1. R. Siedentop et al., "Getting the Most Out of Enzyme Cascades: Strategies to Optimize In Vitro Multi-Enzymatic Reactions," *Catalysts* 2021, 11, 1183., doi: <https://doi.org/10.3390/catal11101183>
2. P. De Santis et al., "The rise of continuous flow biocatalysis – fundamentals, very recent developments and future perspectives," *In React. Chem. Eng.* 5 (12), pp. 2155–2184., doi: <https://doi.org/10.1039/D0RE00335B>
3. D. Medeiros, "DWSIM - Open Source Process Simulator," URL: <https://dwsim.org/>
4. M.J. Menke et al., "Development of an Ontology for Biocatalysis," *Chemie Ingenieur Technik*, 2022, 94: 1827-1835, doi: <https://doi.org/10.1002/cite.202200066>
5. J. Grün et al., "From Coiled Flow Inverter to Stirred Tank Reactor – Bioprocess Development and Ontology Design," *Chemie Ingenieur Technik*, 2022, 94: 852-863., doi: <https://doi.org/10.1002/cite.202100177>
6. J. Range et al., "EnzymeML—a data exchange format for biocatalysis and enzymology," *FEBS J*, 2022, 289: 5864-5874., doi: <https://doi.org/10.1111/febs.16318>
7. S. Lauterbach et al., "EnzymeML: seamless data flow and modeling of enzymatic data," *Nat. Methods*, 2023, 20, 400–402., doi: <https://doi.org/10.1038/s41592-022-01763-1>
8. Jan Range, Frank Bergmann, Johann Rohwer, AnnaReisch, Hannah Dienhart, & SL-2204. (2022). EnzymeML/PyEnzyme: PyEnzyme 1.1.3 (v1.1.3). Zenodo. <https://doi.org/10.5281/zenodo.6457299>
9. S. Arndt et al., "Metadata4Ing: An ontology for describing the generation of research data within a scientific activity". (1.1.0). Zenodo. DOI: <https://doi.org/10.5281/zenodo.770601>
10. J. B. Lamy, "Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies," *Artificial Intelligence in Medicine*. 80., doi: <https://doi.org/10.1016/j.artmed.2017.07.002>

Castellum

A data protection-compliant web application for the subject management of human science studies

Karolina Mader¹ and Maike Kleemeyer¹

¹ Max Planck Institute for Human Development, Germany

Abstract. Research institutions, especially in the human sciences, have been confronted with strict guidelines for the processing of personal data since the European General Data Protection Regulation came into force in 2018. This presents them with new challenges in recruiting and managing study participants and processing the associated data. To meet these challenges, Castellum has been developed at the Max Planck Institute for Human Development since 2016 and has been used successfully since May 2020. Castellum is a turnkey open-source web application for the data protection-compliant management of volunteer data. Among other things, Castellum simplifies study recruitment, appointment management and study implementation. Various institutions have expressed interest in Castellum in the recent past. On the one hand, this may be due to the fact that no comparable open source project exists. On the other hand, Castellum was explicitly designed to be so flexible and expandable that it can be adapted to the workflows and processes of other research institutions with relatively little effort. The use of Castellum has so far been particularly successful for institutions that conduct several studies in parallel, that want to proactively recruit participants from an internal pool of people interested in the study, and that generate data when dealing with these subjects. Since Castellum is subject to the AGPL licence, the software may be used free of charge without restrictions.

Keywords: Castellum, General Data Protection Regulation, Recruiting participants, Open Source

Abstract

At the latest since the European General Data Protection Regulation (GDPR) and the revised Federal Data Protection Act came into force in 2018, strict guidelines apply to the processing of personal data. Research institutions must ensure the implementation of the legal requirements through appropriate technical and organizational measures. This poses new challenges, particularly for research institutions with a focus on human sciences, in the recruitment and administration of study participants and the processing of related personal data. Previous (stand-alone) solutions are often no longer capable of meeting these challenges. However, if the regulations of the GDPR are not complied with, major financial, structural and reputational consequences can occur. At the same time, there is a desire among researchers for a user-friendly, digital application that simplifies the management of subject data and thus efficiently supports the entire research process. However, many institutions lack both the resources and expertise to implement a sustainable and application-oriented solution.

For this reason, Castellum (<https://castellum.mpib.berlin>) has been developed at the Max Planck Institute for Human Development (MPIB) since 2016. This is a turnkey open-

source web application for the data protection-compliant management of subject related data. The development took place in close coordination with the scientific staff and the data protection officer of the Max Planck Society (MPG). Thus, Castellum takes into account relevant aspects of the rules of good scientific practice as well as data security. To ensure that Castellum can be used as an open-source project, it was designed from the beginning to be flexible and expandable so that it can be adapted to the workflows and processes of other research institutions with little effort.

The application provides a clearly defined structure for handling the data of all study participants. Contact information (e.g. name and postal address), recruitment characteristics (e.g. age and education level) and process information (e.g. existing consents and current availability) are stored in Castellum, scientific data is stored outside of it. The focus here is on compliance with provisions of the GDPR and general IT security in dealing with this data. For example, the listed information is strictly separated from each other through an integrated rights and role management. For example, a "recruiter" is only able to view the contact information of potential study participants, but not the study-specific recruitment characteristics. These, in turn, are defined by the "study coordinator" when creating a study, so that only those potential subjects are suggested to the "recruiter" who fit the corresponding study.

In addition, an important function of Castellum is the pseudonym service: This service links scientific data stored outside Castellum with the data in Castellum. This makes it possible to generate a central overview of all the places where the data of subjects has been stored. This in turn helps to efficiently realize the subject rights provided for by the GDPR (e.g. requests for data access and deletion by subjects). Other features of Castellum include booking appointments, storing recruitment and study consents as the legal basis for data retention, and assigning legal representatives.

Castellum has been successfully in productive use at the MPIB since May 2020 and at the Max Planck Institute for Biological Cybernetics since September 2021. However, deficits in the area of subject management have also been identified at other institutions and the need for a comprehensive software solution is seen. Presumably due to the fact that there is no comparable open-source project that covers this broad range of applications, some institutions have already expressed interest in Castellum (e.g. the Universities of Hamburg and Helsinki, the University Medical Center Hamburg-Eppendorf (UKE), Clinic and Polyclinic for Psychiatry and Psychotherapy, Ernst Strüngmann Institute Frankfurt).

That is why we now want to make Castellum (more) known, with the aim of establishing an active application and development community, in line with the slogan of the conference, "Connecting Communities". An intensive (experience) exchange with all users should then serve to continuously improve Castellum and to jointly benefit from best practice approaches and new ideas. Through a broad use of Castellum, changes with regard to data protection regulations only have to be implemented at one central point and all using institutions benefit together. This allows resources to be used much more efficiently.

The technical installation is already supported by detailed documentation. This is available at the following link: <https://git.mpib-berlin.mpg.de/castellum/castellum/-/tree/main/docs/deployment>. In addition, our team is available to advise on questions and suggestions regarding (data) security, compatibility and the technical as well as the organizational use of Castellum and supports pilot projects.

This presentation will give a brief overview of the data protection requirements for the subject management of scientific institutions. Castellum will then be presented as an open-source application and possible solution for implementing these requirements. In particular, the functions and roles covered by Castellum will be described (see Figure 1). Selected functionalities will be demonstrated live. Finally, it will be shown which concrete steps scientific

institutions can take to use Castellum for the first time. Therefore, the contribution is assigned to the programme block "Enabling RDM".

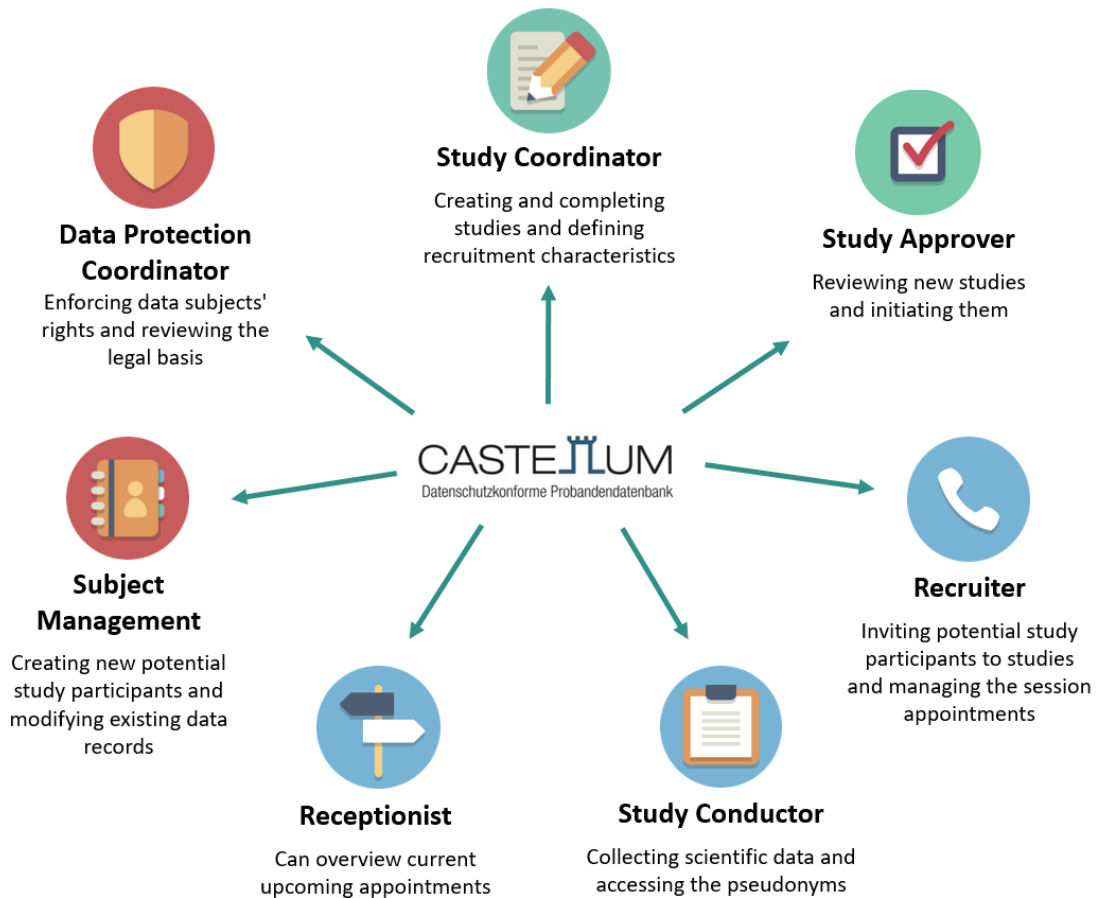


Figure 1. Overview of the role and rights management integrated in Castellum.

Data availability statement

The submission is not based on specific data.

Author contributions

Both authors listed have equally added to the contribution.

Competing interests

The authors declare that they have no competing interests.

FAIRification of Historical Geodata

Automated Metadata Extraction from Archival Maps

Hendrik Herold¹[\[https://orcid.org/0000-0002-2806-3121\]](https://orcid.org/0000-0002-2806-3121), André Hartmann¹[\[https://orcid.org/0000-0003-0384-8501\]](https://orcid.org/0000-0003-0384-8501), Anna Lisa Schwartz²[\[https://orcid.org/0000-0003-0391-3879\]](https://orcid.org/0000-0003-0391-3879), Michael Hellstern², and Markus Schmalz²

¹ Leibniz Institute of Ecological Urban and Regional Development (IOER), Germany

² Bavarian State Archives (GDA), Germany

Abstract. A key challenge for the NFDI community is to share and connect datasets across spatial and temporal scales. In particular for Earth System Sciences (ESS) coupling of different RDIs and access to long-term time series data are crucial. In this paper we propose a workflow pipeline for making analogue bound archival geodata FAIR and hence accessible to the different scientific disciplines and reusable in a long-term perspective. We describe an interface for enabling data exchange between an archive (GDA) and a geospatial RDI (IOER Monitor) as well as other RDIs by applying the FAIR data principles.

Keywords: Enabling RDM, Archival Maps, Geospatial Data, FAIR Principles

1. Motivation and Assets of the RDIs

1.1 Archival Geodata (GDA)

The Generaldirektion der Staatlichen Archive Bayerns (henceforth GDA) is the central state authority for all matters of archiving in Bavaria. The Central State Archive of Bavaria and eight regional state archives are subordinated to GDA as technical and administrative head of the Bavarian State Archives. Part of the technical tasks of GDA is the development and maintaining of central digital infrastructures of the Bavarian State Archives, in particular online-services and the Digital Archive. GDA maintains and strengthens a tight network with national and international partners from different research and science communities and the archival community. In close cooperation with research and infrastructure facilities GDA and their associated archives contribute to the development of sustainable information infrastructures in many ways. Among other initiatives and cooperations GDA develops and maintains a digital archive (DIMAG) according to international standards (ISO 14721, PREMIS 3.0) in cooperation with other German state archives, for recording, high-level long-time content preservation and data provenance and providing of born digitals and develops standardised and automated interfaces for the archiving of born digitals. GDA and the associated archives are long-term-archiving, preserving, describing and providing primary analogue and digital research data and corresponding metadata from the collections of state and non-state archives in Bavaria covering 1200 years from 8th century to the present with the aim of open access and use. This includes hundreds of thousands historical maps containing historical geodata from the late 15th to the 20th century.

1.2 LULC-Monitoring (IOER Monitor)

The Monitor of Settlement and Open Space Development (IOER Monitor, <https://www.ioer-monitor.de/en/>, <http://doi.org/10.17616/R3QF5P>) is a permanent scientific service within the framework of the research-based Political and Social Consultation Service of the Leibniz Institute for Ecological Urban and Regional Development (IOER). This expert information system is intended to assist scientists, administrators, practitioners from business and industry as well as the general public to answer questions regarding ground cover and land use for the whole of the Federal Republic of Germany. It provides base data for the analysis of land use development, particularly regarding the issue of sustainability. The IOER Monitor is an open research data infrastructure (RDI) in Germany providing domain-specific multi-temporal geospatial datasets, services and visualizations for land use and land cover (LULC)-related development of settlements and open space and closely related topics. Its easy-to-use information system provides multi-scale data offers to form a discussion platform that supports spatial development assessment and evidence-based decision making. It contributes to public land-use change discourses by enhancing information offers that can be adopted by other multi-disciplinary data users - even from non-spatial domains. All data and services are freely available. IOER Monitor is committed to offering continuous services implementing FAIR principles (findable, accessible, interoperable and re-usable) and policy-relevant inputs for transformative spatial development [3]. Figure 1 shows the conceptual framework of the RDI. Historic maps are key to fully describing human influence on LULCC and its impact on the Earth System (cf. [1], [2]), however, they demand high efforts in preparing them for digital analysis.

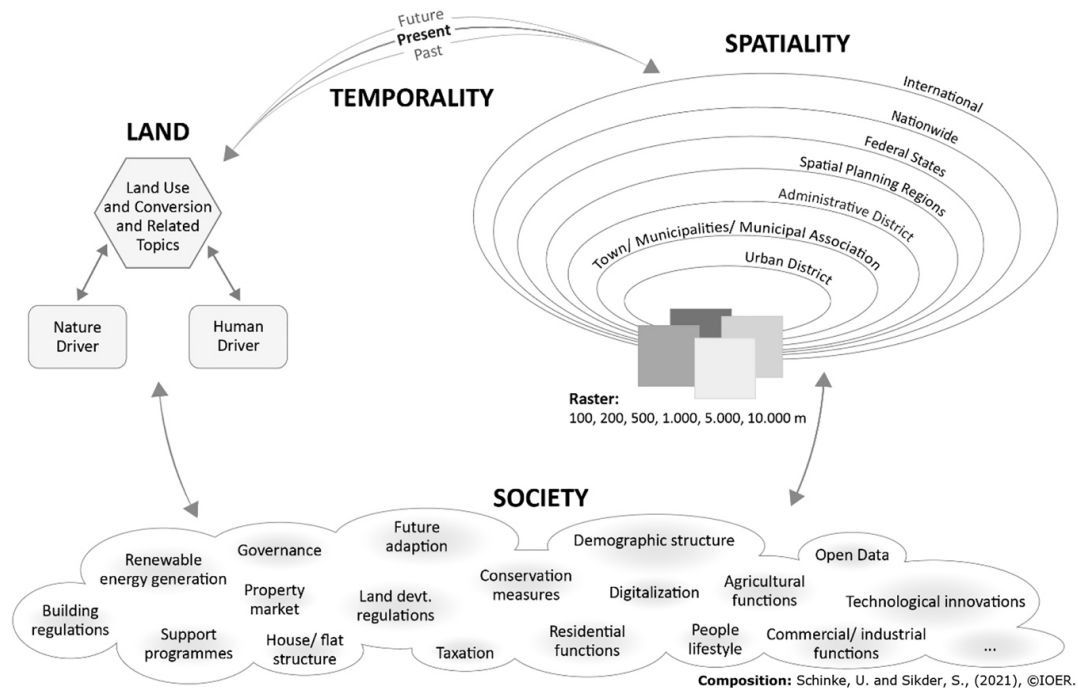


Figure 1. Conceptual framework of the IOER Monitor and the importance of temporality (adopted from [3]).

2. FAIRification of Geodata and Data Life Cycle

2.1 FAIR-Aware Digitization of analogue geospatial archives

Archives and science institutions in Germany and abroad are holding vast quantities of historical maps. For mobilising these analogue geospatial archives the project prepares a best practice workflow starting with Digitization as the first step on the basis of the generalized technical concept of digitization of the Bavarian State Archives. For details see [5] and [6].

2.2 FAIR-Aware Meta-Extraction

We aim at implementing a workflow and data structure, corresponding to FAIR principles (findable, accessible, interoperable and re-usable). Findability-principle is reflected in our general effort to build up a metadata set for archived map scans that can be found, accessed and queried over standard and low hurdle open data pathways. Furthermore, by extracting Map Content, we want to make the information behind, not only the map scan itself, as accessible as possible. This is linked to the more technical part of interoperability, i.e. the use of open standards, formats and software. So by preparing export functionality to a variety of data formats, we allow for a free choice of software on the user side. Finally, the implementation in open source code and software, the definition of an export interface for archiving in addition to the long term archive data provision, enables the reproducibility and reusability of the approach.

Short workflow description (cf. Figure 2):

- QGIS-based Python Toolbox with open source libraries (e.g. OpenCV)
- Map Content Extraction for Meta Data enrichment
- Export of map content and metadata to non-proprietary data formats (e.g. XML, GML, NAS, INSPIRE-conform geodata)

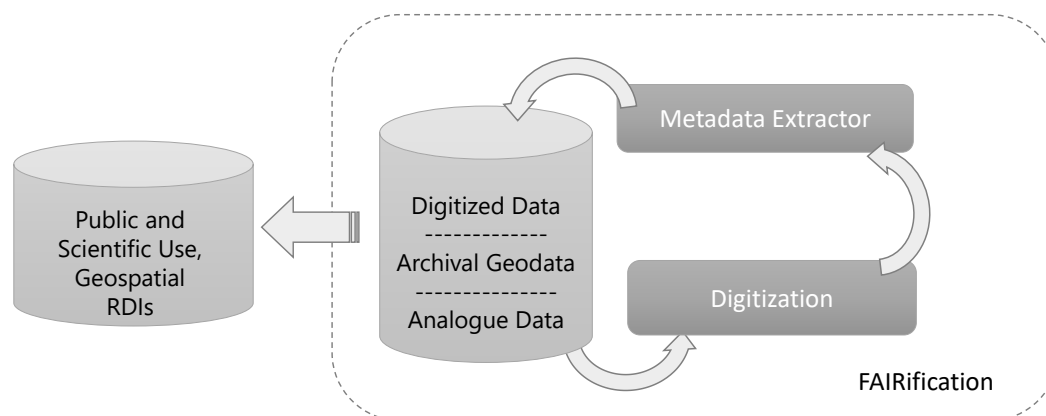


Figure 2. Simplified concept of the proposed pipeline.

2.3 FAIR-Aware Data Life Cycle

The project aims at the FAIRification of historical geodata covering the whole data life cycle. By this means metadata is extracted in interoperable and standardized data and metadata formats and enriched with metadata suitable for current interoperability as well as long term archiving.

For archiving and ensuring the reusability of data the project uses the Generalized XML-Client of GDA addressing the current gap in securing the data life cycle. This allows for automatized structuring, ingesting and description of the data and secures data provenance by documentation of all individual operations as well as the implementation by other repositories. It requires a structured Submission Information Package and a well-documented XSD-file. Within the project archival formats and a list of required data and metadata for documentation for the long-term archiving and reuse of the data are defined as a best practice. Based on the generalized XML-client an interface is aspired for the automated output of all relevant data and metadata.

Data availability statement

The results of this project are metadata sets, which also have a geodata component. The machine-readable textual part of the metadata is relevant for archive and catalogue systems, and will be available likewise. The geodata part of the metadata will be additionally available in a future research data infrastructure. The code and data produced will be made freely available (GitHub, <https://github.com/ioer-dresden>, etc.).

Underlying and related material

The above described RDI IOER Monitor is available here: <http://doi.org/10.17616/R3QF5P>

Author contributions

Hendrik Herold contributed in Funding Acquisition, Conceptualization, Investigation, Data curation, Software, Writing – original draft, Writing – review & editing. André Hartmann contributed in Software, Formal Analysis, Data Curation, Writing – review & editing. Anna Lisa Schwartz contributed in Conceptualization, Data Validation, Writing – review & editing. Michael Hellstern contributed in Conceptualization, Data Validation. Markus Schmalzl contributed in Funding Acquisition, Conceptualization, Investigation, Data curation, Software, Writing – original draft, Writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

The work was partially funded by NFDI4Biodiversity (Flexfund-Project P-06/2022).

Acknowledgement

The authors would like to thank NFDI4Biodiversity and DFG for funding parts of this work.

References

1. H. Herold, "Big Historical Geodata for Urban and Environmental Research", In: Handbook of Big Geospatial Data. Springer, (2021), pp. 475-486, doi: https://doi.org/10.1007/978-3-030-55462-0_18
2. H. Herold, "Geoinformation from the past – computational retrieval and retrospective monitoring of historical land use," Wiesbaden, Springer, 2017, 192 p. <http://dx.doi.org/10.1007/978-3-658-20570-6>
3. G. Meinel, S. Sikder, T. Krüger, „IOER Monitor: A Spatio-Temporal Research Data Infrastructure on Settlement and Open Space Development in Germany”, In: Jahrbücher für Nationalökonomie und Statistik 242, 2022, 1, pp.159-170, doi: <https://doi.org/10.1515/jbnst-2021-0009>
4. Holzapfl, Julian, et al. Quick wins und dicke Bretter. Übernahme und Archivierung von Fachverfahren. Archiv. Theorie & Praxis, vol. 76, no. 1, 2023, pp. 15-24.
5. Puchta, Michael. Automatisierung und Standardisierung. Ein Praxisbericht aus den Staatlichen Archiven Bayerns. Archiv. Theorie & Praxis, vol. 74, no. 3, 2021, pp. 180-186.

6. Puchta, Michael, et al. Fachkonzept für das Digitale Archiv der Staatlichen Archive Bayerns. Staatliche Archive Bayerns, München 2023, doi: <https://doi.org/10.5281/zenodo.7743888>.

Several partners – joint effort: RDM synergies in large scale research

Astrid Schneidewind¹[\[https://orcid.org/0000-0002-7239-9888\]](https://orcid.org/0000-0002-7239-9888), Kilian Schwarz²[\[https://orcid.org/0000-0002-0800-2743\]](https://orcid.org/0000-0002-0800-2743),
Thomas Schörner²[\[https://orcid.org/0000-0002-7213-0352\]](https://orcid.org/0000-0002-7213-0352), Bridget M. Murphy¹[\[https://orcid.org/0000-0002-1354-2381\]](https://orcid.org/0000-0002-1354-2381),
and Martin Erdmann⁴[\[https://orcid.org/0000-0002-1653-1303\]](https://orcid.org/0000-0002-1653-1303)

¹ JCNS at MLZ, FZ Jülich, Germany

² Deutsches Elektronen-Synchrotron DESY, Germany

³ Institut für Experimentelle und Angewandte Physik, Kiel University, Kiel, Germany

⁴ Physikalisches Institut III A, RWTH Aachen, Aachen, Germany

Abstract. We report about the synergies of NFDI, ErUM-Data, and related partners in RDM for large scale research in universe and matter.

Keywords: RDM, infrastructures, synergies

Research at large-scale research infrastructures (RI), combined with external use (user services), does not only require adequate funding, but also dedicated services – including data collection and data management. Scientists in the fields of research with ions, neutrons and photons, of high-energy particle and astro-particle physics, of the physics of hadrons and nuclei, of astronomy and accelerator physics – together called the “ErUM” communities (German for “Research on Universe and Matter”) – therefore join their efforts to meet the challenges of what is commonly referred to as the digital transformation.

These challenges arise on one hand from increasing data rates and volumes – data that need to be transported (e.g. radio astronomy – SKA project), or rapidly analyzed in situ / operando (photon and neutron science) or distributed within a collaboration for independent analysis (e.g. high-energy physics), and the storage of raw data needs to be curated. On the other hand, the dedicated infrastructures limit the storage of raw data. There are in addition aspects of open and FAIR data – often already respected within concepts of international collaboration – and of user-friendliness considering diverse needs, as well as different cultures and traditions across DIG-UM.

Within NFDI two consortia are based with in the “ErUM” communities : PUNCH4NFDI¹ is the consortium of particle, astro-, astroparticle, hadron and nuclear physics, representing scientists from universities, the Max Planck society, the Leibniz Association, and the Helmholtz Association. The goal is to setup a federated and "FAIR" science data platform, offering the infrastructures and interfaces necessary for access to and use of data and computing resources of the involved communities and beyond. DAPHNE4NFDI² is a photons and neutrons (PaN) user-driven consortium which brings scientists from universities and research institutes together with RIs, to develop digital based research data management with the aim of embed-

¹ <https://www.punch4nfdi.de/>

² <https://www.daphne4nfdi.de/english/index.php>

ding FAIR data practices in the community by developing and providing services for data capture, storage, curation and analysis. It serves a wide range of scientific disciplines across physics, chemistry, biology, health and environment sciences partially represented in other NFDI consortia. Although there is a diversity of scientific and technical topics synergy between PUNCH4NFDI and DAPHNE4NFDI include the tradition of long-standing international collaborations and the dependencies and interactions of users and facilities in data collection and delivery and the need to develop artificial intelligence methods.

Base4NFDI as an overarching initiative of all NFDI consortia for the development of basic services supporting the joint efforts of PUNCH and DAPHNE and strengthens the cooperation by coherent work. There is a common interest of both consortia in user-friendliness and compatibility to internationally established standards of basic services to be agreed within the NFDI landscape.

An independent request for a closer data-related cooperation within the ErUM communities is expressed by BMBF, which resulted in the ErUM-Data³ action plan. This referred to the outcome of a BMBF workshop, as a white paper from the ErUM communities entitled "Opportunities of Digital Transformation in Fundamental Research on Universe and Matter"⁴. In a self-organized manner DIG-UM works on five identified topics: Federated infrastructures, User interfaces, Research Data management, Big Data Analytics and Knowledge distribution. Correspondingly five working groups were established in this self-organized manner, to enable the joint scientific communities (DIG-UM) working together⁵. Similar conclusions and working structures were established with the first NFDI consortia. There is, by definition, topical overlap between NFDI and ErUM-Data; it must however be noted that the ErUM-Data comprises a larger community than DAPHNE4NFDI and PUNCH4NFDI.

Still more players exist in the ErUM data landscape with representatives being continuously invited to common discussions and activities:

A number of RIs, experiments and computing infrastructures are operated by the Helmholtz Association (HGF) that in turn, in the form of its centres, contributes to e.g. RDA, NFDI and EOSC. In addition to the established infrastructures, HGF develops its own strategy for exploiting digital techniques to improve its scientific work: the Helmholtz Incubator Framework for Information and Data Science. It contains five platforms: Metadata, AI, Imaging, Federated IT Services, and the Helmholtz information & data Science Academy. Moreover and important for ErUM, within the HGF's research field "Matter", the topic "Data Management and Analysis" (DMA) brings together scientists from facilities either run by Helmholtz centres or operated with Helmholtz involvement with topical experts in computational and data science along with numerous users from diverse communities. The Helmholtz project oriented funding scheme (POF) allows for long-term and strategic support of DMA and its goals.

The combined efforts of the partners described above will be reinforced by data competence centres (DKZ⁶), initialized by BMBF, these will act as sites for data-related education, research and networking, in close cooperation with NFDI. They shall support the cultural change in using data and the resulting data-based innovation.

Given such a manifold of partners with similar challenges and at least partly overlapping actors, it appears quite natural that work on synergies and efficient common structures was started quite early. Educational activities and information channels are already at least partly

³ https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/7/31640_Aktionsplan_ErUM-Data.pdf?__blob=publicationFile&v=7

⁴ http://www.physik.uni-siegen.de/x-ray/publications/erumdata_dina4_30.04.2019_druck.pdf

⁵ <https://erumdatahub.de/en/dig-um/>

⁶ <https://www.bmbf.de/bmbf/shareddocs/bekanntmachungen/de/2022/06/2022-06-21-Bekanntmachung-Datenkompetenzzentren.html>

shared between the various players. Obviously, the ErUM schools (e.g. on Deep Learning – basic as well as Train-the-Trainer) and the expert workshops (e.g. on analysis tools, sustainability in digital transformation) multiplex the network and interaction within the ErUM communities, and create joint knowledge. A special series of workshops on synergies is being established as a forum for information exchange and networking. Within this format, also Helmholtz activities can interact with the NFDI partners: Helmholtz-DMA views both the NFDI and ErUM-Data as natural partners and strives to support common solutions and complementary contributions to the data ecosystem. Beginning collaboration of DAPHNE4NFDI and PUNCH4NFDI with HMC guarantees not only increasing awareness on the importance of metadata, but concurrent metadata standards and agreed formats. In addition, topical connections to other NFDI consortia occur naturally.

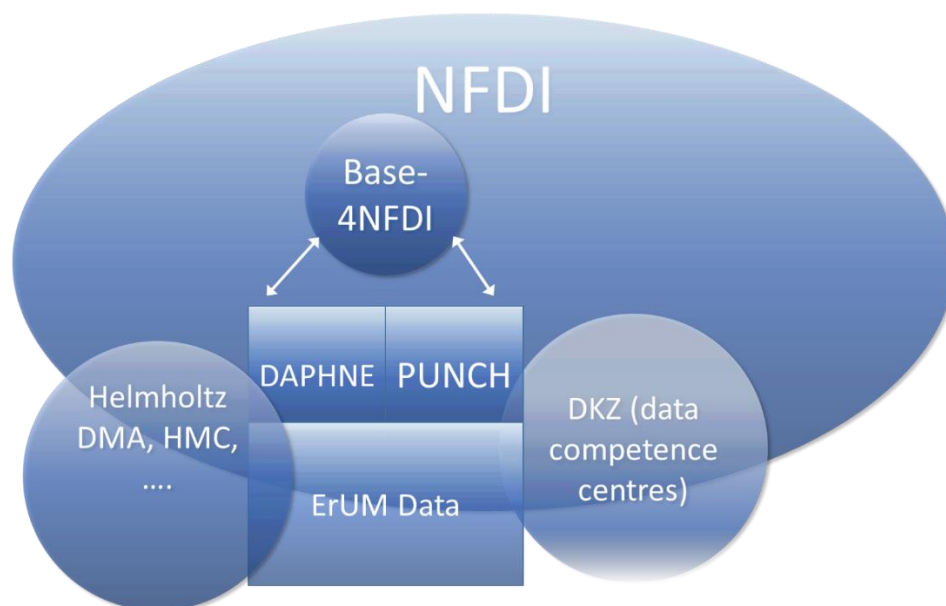


Fig.1: A view point of the NFDI initiatives with their connections to relevant efforts for mastering digital transformation in fundamental and applied research.

The exchange on detailed technical solutions, tools, workflows and strategies is another important aspect of the cooperation between ErUM-Data, NFDI, Helmholtz and others. – The effectiveness of such exchange can e.g. be seen in products like SciCat⁷ or Rucio⁸.

SciCat as a metadata catalogue was created especially for the needs of large and diverse user communities in using data from RIs – the challenge being e.g. to search, find and cite these data even when you are not a member of the original experimental team. University groups, in close contact with the main SciCat developers, now created a reduced SciCat version that allows the handling of data collected in small labs in a way similar to the data from large RIs, in close contact to the main developers.

Rucio provides services and associated libraries for allowing scientific collaborations to manage large volumes of data spread across numerous facilities at multiple institutions and organisations. It was originally developed to meet the requirements of the high-energy physics experiment ATLAS, and now is continuously extended to support the LHC experiments and other diverse scientific communities. DAPHNE4NFDI will test the usability of Rucio for the experiments in photon and neutron science.

Furthermore, joint working groups go for on PIDs, sustainable research software, education and metadata standards.

⁷ <https://scicatproject.github.io/>

⁸ <https://rucio.cern.ch/>

Author contributions

AS wrote the original draft. AS, KS, TSS, BM and ME reviewed and edited the article.

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

We kindly acknowledge the in-kind support of FZJ, DESY, Kiel University and RWTH for the presented work. DAPHNE4NFDI is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 460248799. PUNCH4NFDI is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 460248186. ErUM-Data Hub is funded by BMBF (German Ministry for Education and Research), Förderkennzeichen 05D21PA1.

Data Management Plan Tools: Overview and Evaluation

Carina Becker ¹[\[https://orcid.org/0000-0001-7595-2012\]](https://orcid.org/0000-0001-7595-2012),
Carolin Hundt ²[\[https://orcid.org/0000-0002-7237-965X\]](https://orcid.org/0000-0002-7237-965X),
Claudia Engelhardt ¹[\[https://orcid.org/0000-0002-3391-7638\]](https://orcid.org/0000-0002-3391-7638),
Johannes Sperling ²[\[https://orcid.org/0000-0001-7413-1730\]](https://orcid.org/0000-0001-7413-1730),
Moritz Kurzweil ²[\[https://orcid.org/0000-0003-1413-6673\]](https://orcid.org/0000-0003-1413-6673),
Ralph Müller-Pfefferkorn ¹[\[https://orcid.org/0000-0001-8719-5741\]](https://orcid.org/0000-0001-8719-5741)

¹Center for Information Services and High Performance Computing (ZIH), TU Dresden, Dresden, Germany

²Directorate – Library and Digital Services, Leibniz Institute for the History and Culture of Eastern Europe (GWZO), Leipzig, Germany

Abstract: Data Management Plans (DMPs) are crucial for a structured research data management and often a mandatory part of research proposals. DMP tools support the development of DMPs. Among the variety of tools available, it can be difficult for researchers, data stewards and institutions to choose the one that is most appropriate for their specific needs and context. We evaluated 18 DMP tools according to 31 requirement parameters covering aspects relating to basic functions, technical aspects and user-friendliness. The highest total evaluation scores were reached by Data Stewardship Wizard (703.5), DMPTool (615.5) and RDMO NFDI4Ing (549.5). The tools evaluated satisfied between 10 % and 87 % of the requirement parameters. 11 tools cover at least half of the parameters. In terms of correlation among the tools, which indicates to which degree their scores in the different requirement parameters are alike, we found the highest correlation for ezDMP and GFBio DMPT. Regarding the relatedness between the tools, 85 % of the DMP tools were positively and 16 % negatively correlated. Accounting for the recent developments in the area of DMP tools, this study provides an up-to-date evaluation that can support tool developers in identifying potential improvements, and hosting institutions to select a tool suited to their specific needs.

Keywords: Data Management Plan, DMP Tools, FAIR, Research Data Management

1 Introduction

Data Management Plans (DMPs) are the basis for a structured research data management throughout the data lifecycle [1]. DMPs facilitate especially the administration of data in research groups by describing the joint handling of the data. Additionally, funders require DMPs in research proposals more frequently [2], [3]. There are a variety of tools to support the development of DMPs. From interdisciplinary DMP tools, which are used to write a generic draft DMP, to discipline-specific DMP tools, which support the creation of a DMP in different research fields, such as psychology, biodiversity, engineering, or life sciences. The manual creation of DMPs is very time-consuming, since researchers have to start from scratch and run the risk of not meeting the funder requirements. By using tools, DMPs can be effectively developed and managed. In view

of the large number of offers, the selection of a suitable tool poses a great challenge for researchers. Therefore, it is crucial to analyze these tools [4]. To support the decision of institutions planning to host a DMP tool, this evaluation can also be helpful. Thus, the objectives of this work are as follows:

- A. Identify requirement parameters to evaluate existing DMP tools
- B. Evaluate DMP tools based on the identified parameters
- C. Determine the relatedness between the DMP tools

2 Materials and Methods

We evaluated 18 mainly open access DMP tools (table 1) based on the identified requirement parameters, of which seven provide discipline-specific and eleven generic templates. Eight tools were developed and hosted in Germany. The remaining DMP tools originate from other European countries, USA and Australia. Although DMPTool is based on DMPonline, we consider these tools individually as they differ in their rating scores.

Table 1. Evaluated DMP tools.

DMP tool	Discipline	Hosting/Developers
UWA-DMP	Interdisciplinary	University of Western Australia, Australia
DMP Canvas Generator	Life sciences	Swiss Institute of Bioinformatics, Switzerland
Clarín-d DMP	Humanities/social science	Eberhard Karls Universität Tübingen, Germany
ARIADNE	Archeology	Vast-Lab, Italy
ezDMP	Interdisciplinary	Columbia University, Rutgers University, University of Illinois, USA
GFBio	Biodiversity	GFBio, Germany
TUDD DMP	Interdisciplinary	TU Dresden, Germany
RDMO	Interdisciplinary	Leibniz Institute for Astrophysics Potsdam, University of Applied Sciences Potsdam, Germany
DataWiz	Psychology	Leibniz Institute for Psychology Information, Germany
TUM Workbench	Interdisciplinary	TU München, Germany
QUT	Interdisciplinary	Queensland University of Technology, Australia
ARGOS	Interdisciplinary	OpenAIRE AMKE, EUDAT CDI, Europe
easyDMP	Interdisciplinary	EUDAT, Finland, Norway
NFDI4Plants Dataplan	Plant science	Eberhard Karls Universität Tübingen, Germany
DMPonline	Interdisciplinary	Digital Curation Centre, University of Edinburgh, United Kingdom
RDMO NFDI4Ing	Engineering	University and State Library Darmstadt, Germany
DMPTool	Interdisciplinary	California Digital Library, University of California, USA
Data Stewardship Wizard	Interdisciplinary	Czech Technical University, Dutch Techcentre for Life Sciences, Czech Republic, Netherlands

Based on the findings of 19 expert interviews and a subsequent discussion among the project partners, we identified requirement parameters for the evaluation of existing DMP tools. The parameters were grouped into main categories to show a more detailed view of the rating scores. We focused on the technical requirements in order to identify

DMP tools, which are easy to host and maintain to ensure their adaptability to the specific needs of researchers, institutions, and funders. Furthermore, a weight factor between zero (not relevant) and three (high priority) was assigned to every parameter. For this purpose, the weight factor was determined individually by each member of the research team, and afterwards the arithmetic mean was calculated.

The DMP tools were rated by two different researchers independently according to a fixed rating scheme from zero (poor) to ten (excellent). In a next step, we calculated the arithmetic mean for each requirement parameter. To calculate the final score, the score for each parameter was multiplied by the weight factor. Then, the sum of the rating scores was calculated per main category and for the total score. Furthermore, the percentage of DMP tools with a score greater equal five was calculated. To identify and compare the linear relationships between the tools, that indicates to which degree their scores in the different requirement parameters are alike, the Pearson correlation coefficients based on the scores of the requirement parameters were computed.

3 Results and Discussion

Table 2 shows the 31 identified requirement parameters, which are important for the easy hosting and maintenance of a DMP tool. The parameter 'text modules' is of high importance, since many researchers prefer pre-fabricated text passages, which are automatically generated by the DMP tool. Although such a text might need some refinement by the researchers, it can serve as a first draft of a DMP. The important aspect of machine-actionability [5]–[7] is represented by 'export/import of DMP in tool format' and 'various export formats'.

Figure 1 shows the rating scores of the evaluated DMP tools. The evaluated tools satisfied between 10 % and 87 % of the requirement parameters. 61 % covered at least half of the parameters. The highest total rating scores were attained by Data Stewardship Wizard (DSW) (703.5), DMPTool (615.5) and RDMO NFDI4Ing (549.5). In the main category 'basic functions', DSW also performed best (239.5) followed by DMPonline (220) and EasyDMP (205). The three best performing tools in terms of 'technical aspects' were again DSW (190), DMPTool (190) and RDMO NFDI4Ing (181). The most user-friendly ones were DSW (274), DMPTool (225.5) and RDMO QUT (220). Comparing the results of our study with those of Gajbe et al. [4], who analyzed 14 tools, there are quite different results in the evaluations. 72 % of the evaluated tools provided source code, compared to 64 % in Gajbe's study. The majority (79 %) of tools from Gajbe were open access, however our findings showed 100 % open access. 67 % of our analysed tools provided user-friendly guidance, compared to 86 % of Gajbe's results. Most of the tools evaluated by Gajbe (86 %) had a user guide, while our results could confirm this for only 44% of the tools. In Gajbe's study, 64 % provided an option to share the DMP with others, whereas our results showed 56 %. We found that 67% of the DMP tools provided more than one export format, compared to 57 % in Gajbe's results. Pre-formulated filterable answer options were supplied by 64 % of the tools in Gajbe's study, while our study resulted in 56 %. In both studies, all tools provided open text fields.

Concerning the relatedness, 85 % of the DMP tools showed a positive and 16 % a negative correlation. The Pearson correlation of the tools (figure 2) is highest for ezDMP and GFBio (0.9). We did not find such a strong correlation as Gajbe et al. for DMPonline and DMPTool (1), but only a correlation coefficient of 0.7. The correlations for the other tool pairs differ from Gajbe as well.

The differences between our and Gajbe's results might be explained by the different tools evaluated, although eight tools were the same. Furthermore, the tool properties have changed over the years between Gajbe's analysis (2020) and our study (2023). The individual way of evaluating the tools could have been divergent as well.

Table 2. Requirement parameters are grouped into main categories (dark gray) and subcategories (light gray). Priority 3 = high, priority 0 = not relevant. DSGVO - General Data Protection Regulation.

Parameter	Priority
1 BASIC FUNCTIONS	
1.1 Access	
1.1.1 Open access with login	3
1.1.2 Open access without login	3
1.1.3 Encryption	2
1.1.4 DSGVO compatibility	3
1.2 Storage and Export	
1.2.1 Saving	3
1.2.2 Export/import of DMP in tool format	3
1.2.3 Various export formats	3
1.3 Collaboration	
1.3.1 Share DMP with collaborators	3
1.3.2 Track changes	1
1.3.3 Commenting function	2
1.3.4 Control levels	2
2 TECHNICAL ASPECTS	
2.1 Editing	
2.1.1 Editor access (CMS with roles)	3
2.1.2 Modularity ('generic' and 'institution specific')	3
2.1.3 Frontend/backend access	2
2.1.4 Easy maintenance of content	3
2.1.5 Sustainability of the software (updates and development)	3
2.2 Transparency	
2.2.1 Open source	3
2.2.2 FAIRness	2
3 USER FRIENDLINESS	
3.1 Assistance	
3.1.1 User friendly guidance	3
3.1.2 Pre-formulated filterable answer options	3
3.1.3 Text modules	3
3.1.4 Text sections (short DMP)	3
3.1.5 Preview of text modules (what you see is what you get)	2
3.1.6 User guide	3
3.1.7 User feedback	2
3.2 Design/Structure	
3.2.1 Layout/usability	3
3.2.2 Progress	2
3.2.3 Breadcrumbs (navigation)	2
3.2.4 Highlighting unanswered questions	3
3.2.5 Skipping questions	3
3.2.6 Open text fields	3

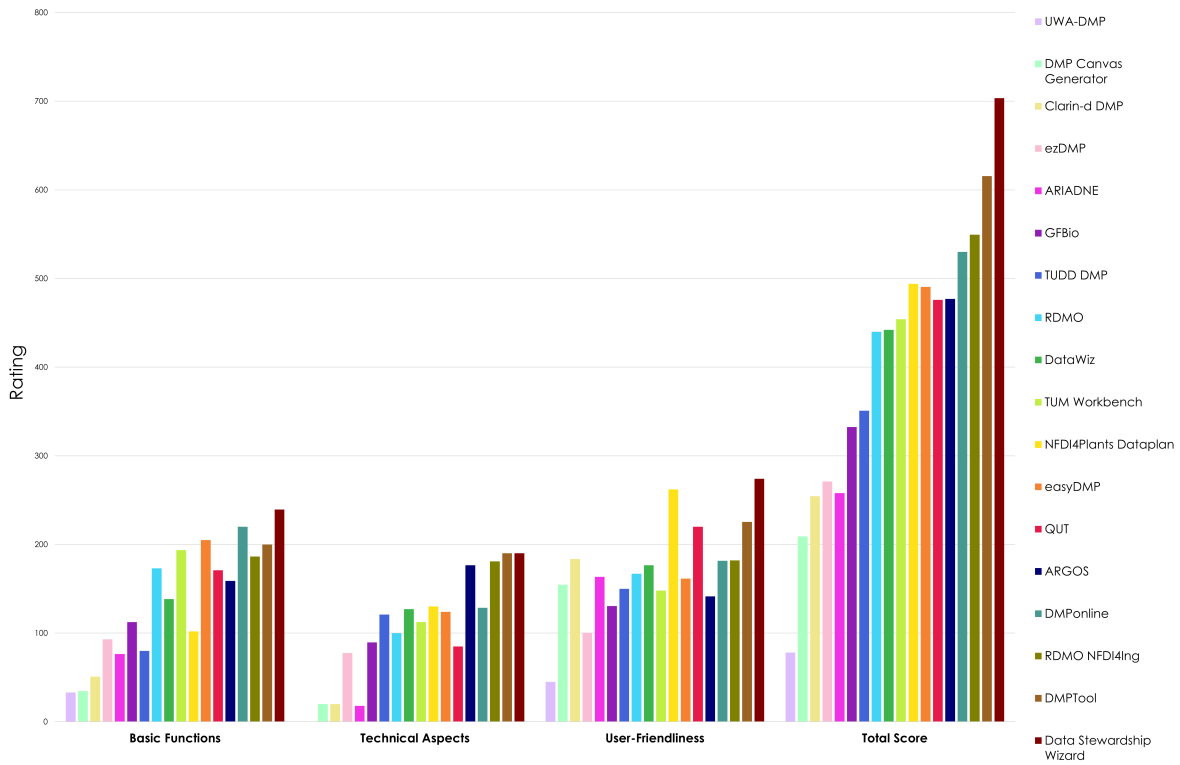


Figure 1. Evaluation results of 18 DMP tools. The rating scores are grouped by main categories (x-axis).

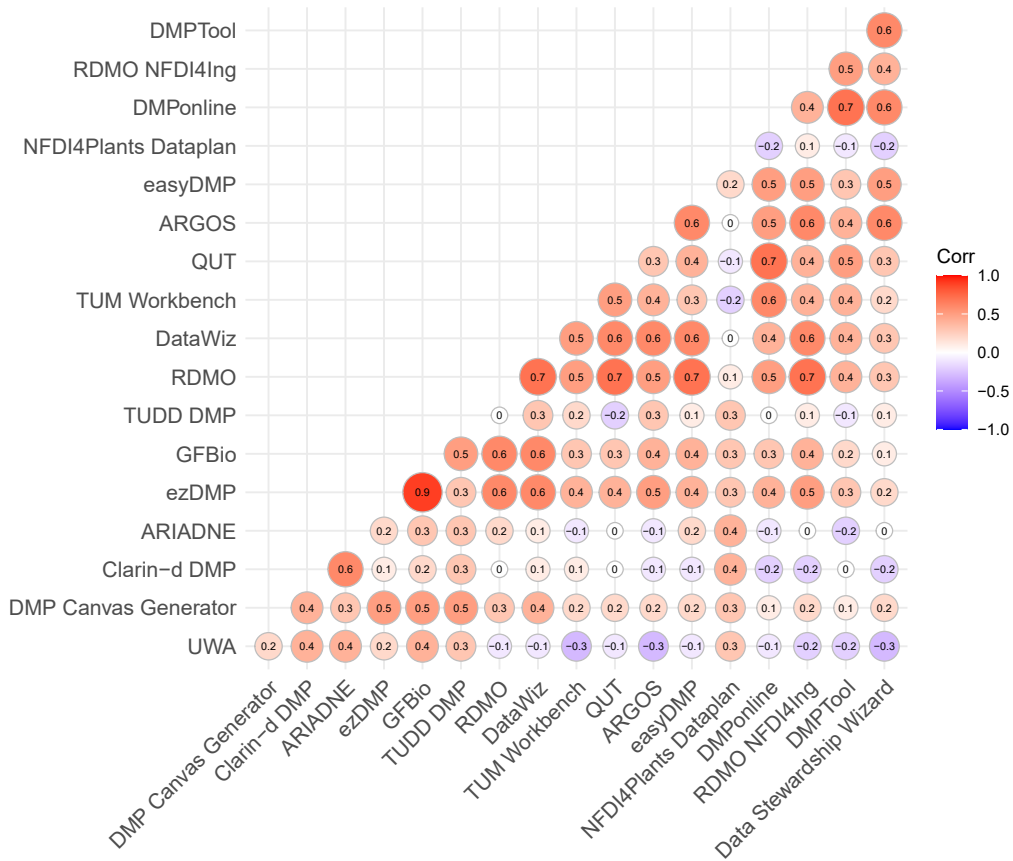


Figure 2. Correlation of 18 DMP tools based on the requirement parameters.

4 Conclusion

Our results show that Data Stewardship Wizard, DMPTool and RDMO NFDI4Ing are the highest rated DMP tools and can be recommended for researchers and institutions as flexible tools for hosting. In the light of recent developments in the area of DMP tools, this study provides an up-to-date evaluation of 18 DMP tools according to 31 parameters covering basic functions, DMP contents, technical aspects and user-friendliness. The results can support tool developers to identify potential improvements and hosting institutions to select a tool suited to their specific needs.

Data availability statement

The data supporting the results of our contribution can be accessed once the more extensive work is published in a peer reviewed journal.

Underlying and related material

There is no supplementary material.

Author contributions

Carina Becker: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, writing - original draft

Carolin Hundt: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, investigation, writing - review and editing

Claudia Engelhardt: investigation, writing - review and editing

Johannes Sperling: investigation

Moritz Kurzweil: investigation, writing - review and editing, funding acquisition

Ralph Müller-Pfefferkorn: conceptualization, investigation, funding acquisition, methodology, supervision

Competing interests

The authors declare that they have no competing interests.

Funding

This study is co-funded with taxes based on the budget passed by the Saxon state parliament (research proposal number 100607005).

References

- [1] A. Ball, *Review of Data Management Lifecycle Models (version 1.0)*. Citeseer, 2012.
- [2] W. K. Michener, "Ten simple rules for creating a good data management plan," *PLoS computational biology*, vol. 11, no. 10, e1004525, 2015. DOI: <https://doi.org/10.1371/journal.pcbi.1004525>.

- [3] European Research Council and Scientific Council, *Open research data and data management plans - information for erc grantees*, https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf, Accessed: 19.04.2023, 2022.
- [4] S. B. Gajbe, A. Tiwari, R. K. Singh, *et al.*, "Evaluation and analysis of data management plan tools: A parametric approach," *Information Processing & Management*, vol. 58, no. 3, p. 102480, 2021. DOI: <https://doi.org/10.1016/j.ipm.2020.102480>.
- [5] T. Miksa, P. Walk, P. Neish, *et al.*, "Application profile for machine-actionable data management plans," *Data Science Journal*, vol. 20, no. 1, 2021. DOI: <https://doi.org/10.5334/dsj-2021-032>.
- [6] T. Miksa, S. Oblasser, and A. Rauber, "Automating research data management using machine-actionable data management plans," *ACM Transactions on Management Information Systems (TMIS)*, vol. 13, no. 2, pp. 1–22, 2021. DOI: <https://doi.org/10.1145/3490396>.
- [7] N.-M. Pham, H. Moulaison-Sandy, B. W. Bishop, and H. Gunderman, "Data management plans: Implications for automated analyses," *Data Science Journal*, vol. 22, no. 1, 2023. DOI: <https://doi.org/10.5334/dsj-2023-002>.

When Data Crosses Borders – Join Forces!

Multidisciplinary Use Cases Within NFDI

Barbara Ebert¹[\[https://orcid.org/0000-0003-3328-6693\]](https://orcid.org/0000-0003-3328-6693), Sami Domisch²[\[https://orcid.org/0000-0002-8127-9335\]](https://orcid.org/0000-0002-8127-9335), Christin Henzen³[\[https://orcid.org/0000-0002-5181-4368\]](https://orcid.org/0000-0002-5181-4368), Jimena Linares¹[\[https://orcid.org/0000-0003-3847-2663\]](https://orcid.org/0000-0003-3847-2663), Kati Mozygemba⁴[\[https://orcid.org/0000-0002-0326-1607\]](https://orcid.org/0000-0002-0326-1607), Bernhard Miller⁵[\[https://orcid.org/0000-0002-4385-7245\]](https://orcid.org/0000-0002-4385-7245), Bernhard Seeger⁶[\[https://orcid.org/0000-0002-9362-153X\]](https://orcid.org/0000-0002-9362-153X), Jörg Seegert³[\[https://orcid.org/0000-0001-9357-2830\]](https://orcid.org/0000-0001-9357-2830)

1 GF Bio e.V.

2 Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin

3 Technische Universität Dresden

4 Universität Bremen, FDZ Qualiservice

5 GESIS - Leibniz-Institute for the Social Sciences, Mannheim and Cologne

6 Philipps-Universität Marburg

Abstract. A multidisciplinary use-case for integrating research data involves data, software and methods across research disciplines to address research questions that cannot be effectively addressed by a single discipline or method alone. Multidisciplinary use-cases thus contribute to NFDI's goal to make a significant contribution to answering novel interdisciplinary research questions. This paper presents some selected examples of multidisciplinary use-cases from different NFDI domains that illustrate the efforts needed to reap the additional benefits of these use-cases. It concludes by proposing cornerstones for the research data management of multidisciplinary use-cases.

Keywords: NFDI, research data management, use-cases, collaboration, interdisciplinary research, multidisciplinary research data

1. The Relevance of Multidisciplinary Use-Cases

1.1 Substantive Relevance and Goals of NFDI

Multidisciplinary use-cases for integrating research data involve data, software and methods across research disciplines to address research questions that cannot be effectively addressed by a single discipline or method alone. These use-cases require collaboration across multiple research disciplines and likewise the integration of heterogeneous data, e.g., experimental, observational, and computational data, and management as well as processing workflows. They also require communication about a common understanding of concepts and methods (e.g., the understanding of "interview" can differ between disciplines and methodological approaches). Multidisciplinary use-cases thus help developing a more comprehensive and nuanced understanding of complex phenomena, by leveraging the complementary strengths across research disciplines and analytic approaches including software and workflows, as well as data types. This is reinforced in the German research system by the goals underpinning NFDI: "NFDI will also make a significant contribution to answering novel interdisciplinary research questions with a high societal impact." [1, p.2] This is strongly tied to the German Council of Information Infrastructures' consideration that "perspectives for the development of transformed or enriched data sets or "data products" should be conceived to create opportunities for wider re-use in other disciplines, fields, and domains." [2, p.3]

1.2 Topic and data-driven perspectives

In principle, there are two perspectives to identifying multidisciplinary use-cases: one is based on research topics, the other is data-driven. The first approach is driven by the desire to tackle complex questions with scientific rigor. This strengthens the need to link data from different methodological approaches and disciplines to solve complex research questions. Examples include estimating the socio-economic impacts of climate change and biodiversity loss or assessment of conservation efforts. Climate change's overarching relevance as well as its complexity have brought, among others, researchers from the Earth System Sciences (ESS), the life sciences and the social sciences to join forces and approach the same questions from different angles. To do so it is necessary to create ways for the various types of data to be combined and used for different forms of analysis.

Likewise, disaster prevention and risk management resp. and esp. its aspects of vulnerability, preparedness and resilience also require a multidisciplinary approach, e.g., by linking georeferenced survey data with existing crisis-relevant contextual information of various disciplines.

The second approach is not motivated by known substantive research questions but seeks to enable the identification of relevant patterns through enhancing the data. A data-driven approach as envisioned in [3] will thus emphasize the machine-actionability of data from different sources and this facilitates their interoperability. Researchers – and artificial intelligence – will then, for example, be able to look for patterns – such as health effects (measured e.g., through treatment data) in brightly lit areas (measured using satellite data).

To be sure, the first approach also benefits from richly annotated, machine-actionable metadata.

2. Benefits of Multidisciplinary Use-Cases and Examples

Data drives a better understanding of complex issues. Yet multidisciplinary use-cases generate particular benefits: 1) data can be contextualized in the knowledge of each participating field, which 2) given the likely relatively unconnected nature of research in the area offers new hypotheses that 3) can be tested by methods from each of the participating disciplines and

thus 4) will increase knowledge in likely more than one field. Multidisciplinary use cases therefore promise non-linear benefits as compared to single-discipline use-cases. The examples below illustrate that, in order to reap these benefits, some effort is required.

2.1 The *hydrographr* R-package

Freshwater ecosystems are relevant to both Earth Systems Sciences and especially to Hydro Sciences and - as a habitat - to biodiversity research. Data-wise they are uniquely characterized by their longitudinal, i.e., the up- and downstream connectivity between water bodies. Yet, this connectivity is largely neglected in spatial freshwater analyses on e.g., biodiversity. Since small streams contribute by ca. 70% to the entire stream network length, a sufficiently high resolution is required resulting in large amounts of data. Any tool should therefore (i) efficiently harness the large amount of data, while (ii) allowing users to stay in their common environment, i.e., R.

The *hydrographr* R-package [4] was developed within a 4-month use case in NFDI4Biodiversity - on the basis of a pilot project of NFDI4Earth - that aims to lower the burden for potential users by offering easy-to-use functions within R. The package is tailored towards the high-resolution Hydrography90m [5] stream network data and capitalizes on fast and RAM-efficient open-source command-line GIS software (Geospatial Data Abstraction Library/ OpenGIS Simple Features Reference Implementation (GDAL/OGR), GRASS GIS [6] and AWK) within the Linux and Windows environments. Regardless of the operating system, users can hence create their workflows in R on their local computers employing high-resolution network data across large spatial extents, since the actual data processing is taking place outside R. Functions include wrappers to download, process, read and extract data, as well as to assess network distances and perform connectivity analyses. The package is hosted on GitHub and is available at <https://github.com/glowabio/hydrographr>.

2.2. FEdA BiodiWert Data

FEdA [7] was launched to fund interdisciplinary projects on the impact of nature conservation efforts in Germany. Projects are selected in annual calls. FEdA entered into cooperation on research data management with NFDI4Biodiversity. Workshops on FEdA's first call for proposals in FEdA's BiodiWert (valuation and preservation of biodiversity in politics, economy, and society) served to discuss data management plans (DMPs) and link up with domain-specific services. Most of the BiodiWert projects have multi-method research approaches, like the SLInBio project (Urban Lifestyles and the Valuation of Biodiversity – Dragonflies, Grasshoppers, Bumblebees & Co) [8] referring to biodiversity and environmental data (occurrence, molecular, ecotoxicological), geographical data, and empirical social research data (interviews, questionnaires, surveys, and documents), among other data.

To support this type of interdisciplinary project, a workflow was suggested across data centers from NFDI4Biodiversity and KonsortSWD using the Helpdesk service of the NFDI4Biodiversity's partner GFBio [9]. In the back office, experts and curators from GFBio's Data Centers [10] and experts from the KonsortSWD partners such as the Research Data Centers Qualiservice [11] and GESIS assist the creation of DMPs and prepare specific data types for archiving based on relevant community standards. Together with the FEdA coordination office, a FEdA data policy [12] was developed to create a guideline for the archiving and sustainable publication of research data.

The collaboration with the FEdA initiative constitutes a pilot project for consulting on multidisciplinary data management under NFDI's umbrella. For the partners involved, it required to create, re-think workflows, methodologies, and also enhanced the understanding of the different scientific practices in social sciences and biology.

3. The Way Ahead

Overall, the value of multidisciplinary use-cases for data sharing lies in their ability to facilitate collaboration between different scientific disciplines and methodological approaches. By leveraging heterogeneous data sources, researchers can gain new insights and make more accurate predictions about complex phenomena. Additionally, by supporting collaborations across different consortia, the NFDI can help break down barriers between research disciplines and promote a more holistic approach to science

In a first summary, we identified the following important questions for successfully detecting and processing multidisciplinary use cases across research data infrastructures.

1. How can we find meaningful multidisciplinary use-cases?
2. How can the needs of researchers, the various discipline-specific issues, and scientific methods be taken into account and integrated?
3. What are the ways and processes to systematically explore and exploit use-cases?
4. How can we encourage researchers working on multidisciplinary use-cases in the future?
5. What are the specific requirements in multidisciplinary use-cases? Are there limitations? Which obstacles have to be overcome for a good scientific service?
6. What challenges and opportunities arise from the consideration of discipline-specific or multidisciplinary use-cases for the development of (national) research data infrastructures?
7. How can multidisciplinary use-cases be stimulated or initiated?
8. How do we consider legal and ethical problems that occur by combining different data types?

Data availability statement

not applicable.

Underlying and related material

not applicable.

Author contributions

All authors have contributed to the **conceptualization** and the **writing – original draft**.

Competing interests

The authors declare that they have no competing interests.

Funding

This contribution is funded by grants from the German Research Foundation (DFG) within the framework of the agreement between the Federal Government and the Länder on the establishment and funding of the National Research Data Infrastructure (NFDI) of 26 November 2018:

KonsortSWD – grant no. 442494171

NFDI4Earth – grant no. 460036893

NFDI4Biodiversity – grant no. 442032008

Moreover, we acknowledge funding by the Leibniz Competition (J45/2018) to SD.

Acknowledgement

-

References

1. Gemeinsame Wissenschaftskonferenz von Bund und Ländern, Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018, November, 2018, <https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/NFDI.pdf>
2. RfII – German Council for Scientific Information Infrastructures, Discussion Paper on the Enhancement of Research Data Infrastructures, 2020, <https://rfii.de/?p=4422>
3. Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). A. J. Hey (Ed.). Redmond, WA: Microsoft research.
4. Üblacker, M. M., Grigoropoulou, A., Garcia Marquez, J., Torres Cambas, Y., Schürz, C., Flourey, M., Tomiczek, T., Bremerich, V., Amatulli, G., & Domisch, S. (2023). hydrographr: Scalable Hydrographic Data Processing in R. R-package available at <https://glowabio.github.io/hydrographr/>.
5. Amatulli, G., Marquez, J., G., Sethi, T., Kiesel, J., Grigoropoulou, A., Üblacker, M. M., Longzhu, Q., S., Domisch, S. (2022). Hydrography90m: a new high-resolution global hydrographic dataset. Vol 14 (10), <https://doi.org/10.5194/essd-14-4525-2022>.
6. Neteler, M., Bowman, M., H., Landa, M., Metz, M. (2012). GRASS GIS: A multi-purpose open source GIS, Environmental Modelling & Software, Vol 31, pp. 124-130. <https://www.sciencedirect.com/science/article/abs/pii/S1364815211002775>
7. Linares, J., Ebert, B., Eberhardt, J., Frohne, K., Sauerland, K., Mozygemba, K., Miller, B., Collaborative Research Data Management Support for FEdA projects. Zenodo. <https://doi.org/10.5281/zenodo.7624583>.
8. SLInBio - Urban Lifestyles and the Valuation of Biodiversity – Dragonflies, Grasshoppers, Bumblebees & Co. FEdA - Biodiversity Initiative for the Conservation of Biodiversity. <https://www.feda.bio/en/slinbio/> (accessed on 23.04.2023)
9. GFBio - German federation for Biological Data. <https://www.gfbio.org/> (accessed on 23.04.2023)
10. Data Centers GFBio - German federation for Biological Data. <https://www.gfbio.org/data-centers/>. (accessed on 23.04.2023)
11. Qualiservice. <https://www.qualiservice.org/en/>. (accessed on 23.04.2023)
12. Taffner, J., Eberhardt, J., "FEdA Forschungsdaten-Policy" Zenodo. <https://zenodo.org/record/7798305>.

Science Data Platform and Digital Research Product

Harry Enke¹ and Thomas Schörner-Sadenius²[\[https://orcid.org/0000-0002-7213-0352\]](https://orcid.org/0000-0002-7213-0352)

¹ Leibniz-Institute for Astrophysics Potsdam, Germany

² Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany

Abstract. The participating scientific communities in PUNCH4NFDI have long experiences with data collections which are often huge, compared to other NFDI communities, and also – from necessity – organized collective access to these data collections. The focus of the work, next to the questions of determining the common ground for identifying and finding the data across the existing community silos, is to make use of the vast experiences for operating collaborative execution environments for large distributed communities. Leaving the always implied questions of AAI and Identity Management aside, building and operating a distributed and heterogeneous execution environment is only feasible using an approach which borrows heavily from the notion of micro-services, with a few building blocks which serve as connecting elements.

1. Overview

There is the required capability to enable code execution near the prepared (previously assembled) data set, or – alternatively - temporary location, where simulated or other processed data can be placed and later be transported to a more permanent storage. There are solutions for this caching problem. There are also solutions for the transmission of executable files as containers to remote computing resources.. One task area of PUNCH4NFDI is dedicated to make this operational for the PUNCH4NFDI Science Data Platform (SDP).

Another crucial element is a facility for customising and building the appropriate code execution environment within a container. The containers need a registry for access in a workflow, either by submission to customized queues or to a workflow engine. The workflow engine allows chaining several steps of a workflow, starting from data collection. allocation of the data in a cache, calling the prepared container using compute resources within the platform, finally retrieving and storing results for further processing. Gitlab (or similar products) provide the tools to build, register and deliver such containers, manage the software and allow to work collaboratively in this process. A workflow engine, REANA, supporting different languages (CWL, Snakemake, simple serial and parallel workflow specifications) is also available. These elements need to be orchestrated, which is feasible within the cope of PUNCH4NFDI as the elements themselves are already developed by community providers.

But this alone does not constitute a suitable environment for users. Working on a scientific topic takes often weeks or month, so a means to store the status, the intermediate results etc. in a convenient manner led to the notion of the Digital Research Product. This DRP uses available descriptors for data, software etc., wraps this information into a package and stores the package in a database. Packing and unpacking is done within the portal of the SDP, and this provides a comfortable working environment, especially if data from different sources of different communities with only partially compatible data structures are to be processed. The DRP was conceived by acknowledging, that to ‘integrate’ the data, even only from fundamental

physics, e. g. by one new metadata schema is not feasible. Mapping the scientific understanding is only partially possible. PUNCH4NFDI is building bridges between different realms and disciplines, where often data and metadata are mixed in different ways, partially using also specialized file systems and software, which carry contextual knowledge.

Data availability statement

-

Underlying and related material

-

Author contributions

-

Competing interests

The authors declare to not having competing interests

Funding

-

Acknowledgement

-

References

-

CorWiz a Platform for Exploring Corrosion Data and Accessing Corrosion Models

Sven Berger¹[\[https://orcid.org/0000-0001-6083-7038\]](https://orcid.org/0000-0001-6083-7038), Aravinth Ravikumar¹[\[https://orcid.org/0000-0003-3003-2700\]](https://orcid.org/0000-0003-3003-2700), Mikhail Zheludkevich¹, and Daniel Hoeche¹[\[https://orcid.org/0000-0002-7719-6684\]](https://orcid.org/0000-0002-7719-6684)

¹ Helmholtz-Zentrum hereon GmbH, Geesthacht

Abstract. Corrosion is a major cause of material degradation and failure in various industries and applications. Damage caused by corrosion causes billions in damage each year and is a major cause of infrastructure degradation. Besides that, corrosion mechanisms, countermeasures and effects are generally not good understood. The platform **CorWiz** will provide an easily accessible way to corrosion research and data.

Research data on corrosion mechanisms, rates, and prevention methods are essential for developing effective solutions and improving the performance and reliability of materials. However, corrosion research data are often scattered, inconsistent, or inaccessible, limiting their reuse and impact. Providing a more straightforward way to access available data and models has thus a significant impact on the research field and application areas.

In the following we present the components of the minimum viable platform focusing on stainless steel corrosion.

Keywords: Corrosion, Platform, Data-management, Corrosion-Modeling, Stainless-Steel corrosion

1. Introduction

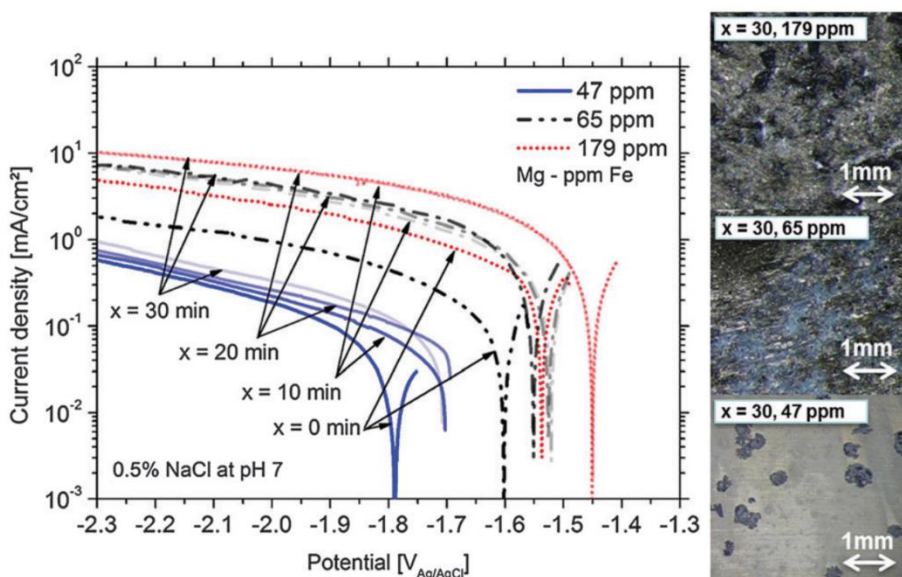


Figure 1: Example corrosion data from measurement [1]

Stainless steel is one of the most used materials in industrial applications due to its excellent corrosion resistance properties. However, stainless steel can still be susceptible to corrosion under certain conditions. Corrosion in stainless steel occurs when the protective layer of chromium oxide on the surface of the metal is damaged or removed, allowing oxygen and moisture to interact with the underlying metal. This can result in various forms of corrosion, such as pitting, crevice corrosion, or stress corrosion cracking, which can compromise the integrity of the material and impact its performance. Therefore, it is important to understand the causes and prevention methods of stainless-steel corrosion to ensure its long-term durability and reliability.

The envisioned platform builds upon existing technologies including Kadi4mat, Plotly and others to provide access to corrosion data and models using an intuitive web interface. While allowing the data to be used interoperable through a consistent metadata schema. Available data can be freely downloaded from the repository, while the platform can also be used to analyze data on the platform for processing and visualization.

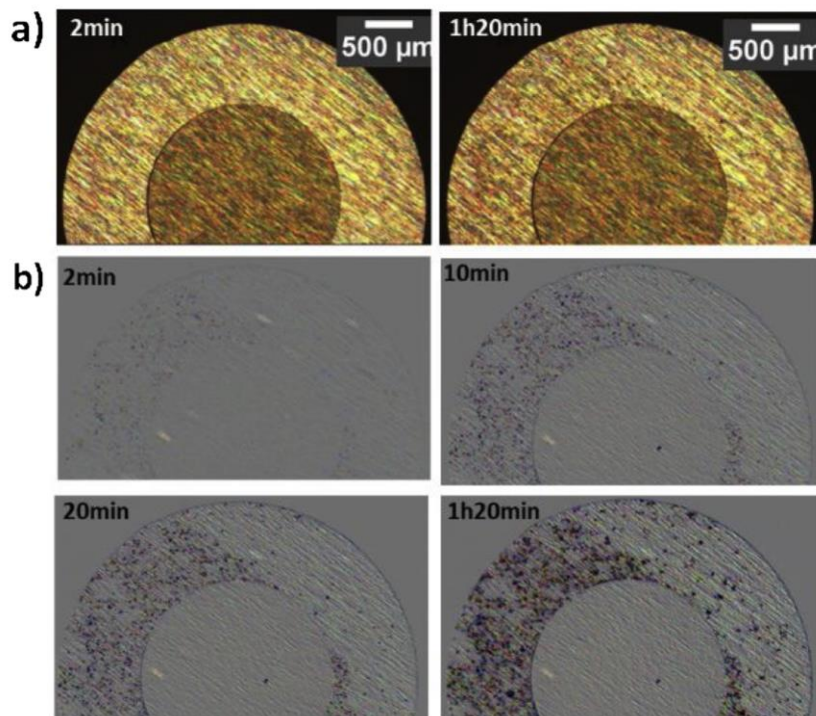


Figure 2: Optical images (a), and difference images by DVIT (b) over AA2024-Ti sample during immersion in 0.05 M NaCl. [2]

2. Data sources

Corrosion research data can be obtained from various sources depending on the objectives and scope of the study. And vary widely in the form of data ranging from electrostatic measurements (see Fig. 1), optical measurements (see Fig. 2) over to simulations (see Fig. 3). Some of the common data sources are:

- Laboratory experiments: These involve conducting controlled tests on materials or structures under simulated environmental conditions to measure their corrosion rates, mechanisms, and effects. The type of data obtained from laboratory experiments can include electrochemical measurements, weight loss measurements, surface analysis techniques, mechanical testing, etc.

- Field studies: These involve monitoring and evaluating the performance of materials or structures exposed to natural or service environments over a period of time. The type of data obtained from field studies can include visual inspection, non-destructive testing, corrosion coupons, sensors and probes, etc.
- Literature review: This involves collecting and analyzing existing information from published sources such as journals, books, reports, standards, etc. The type of data obtained from literature review can include theoretical models, empirical equations, statistical analysis, case studies, best practices, etc.
- Data repositories: These involve accessing and using online databases that store and share corrosion research data from various sources and domains. The type of data obtained from data repositories can include metadata (such as authorship, date, location), raw data (such as numerical values), processed data (such as graphs), derived data (such as indicators), etc.

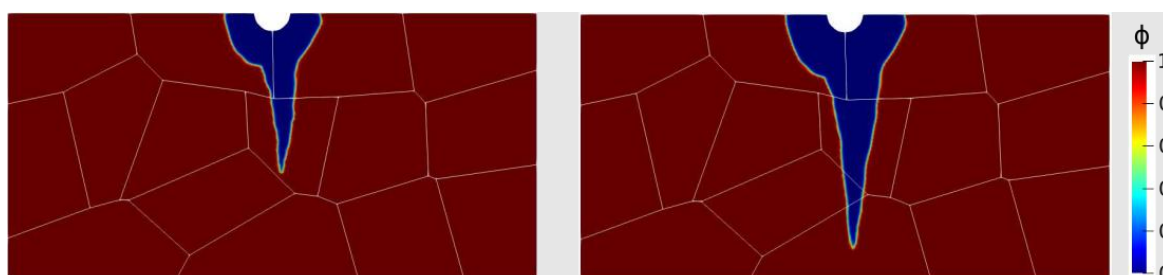


Figure 3: Evolution of the phase field in a pitting corrosion process of aluminium. [3]

3. CorWiz

CorWiz (see Fig. 4) is a platform that helps engineers and researchers access public corrosion data. Users can explore corrosion data using various filters, charts, and maps, and gain insights into the corrosion behavior of different materials and environments. Moreover, the web application enables users to explore and use corrosion models that can predict the corrosion rate, life cycle cost, and risk of failure of various structures and components. In the first phase of **CorWiz** we will limit ourselves to stainless steels and lab measurements of mass loss due to corrosion.

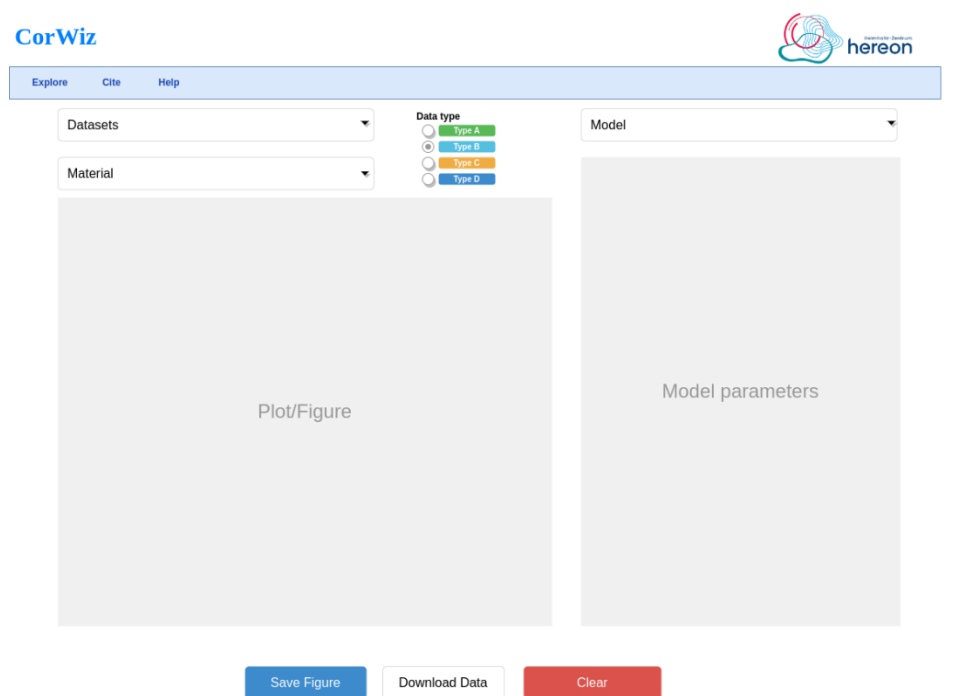


Figure 4: Mock Up of the Web Interface

4. Backend

Kadi4mat is a software tool that allows users to analyze and visualize corrosion data from various sources. It can handle different types of corrosion measurements, such as electrochemical impedance spectroscopy (EIS), potentiodynamic polarization (PDP), and weight loss (WL). Kadi4mat also provides features for data preprocessing, curve fitting, parameter extraction, and statistical analysis. Kadi4mat is used to store and process data to be presented on the **CorWiz**.

5. Conclusion

The minimal viable platform implemented during this first phase of **CorWiz** will demonstrate the impact of Research data management (RDM) and application of FAIR principles in corrosion research. Which will allow us to build upon and extend to further materials, more complex models and workflows.

Competing Interests

The authors declare that they have no competing interests.

Funding

CorWiz is funded by a NFDI seed fund.

References

1. D. Hoeche et al., "The effect of iron re-deposition on the corrosion of impurity-containing magnesium", *Phys. Chem. Chem. Phys.*, 2016, 18, 1279, doi: 10.1039/C5CP05577F.
2. D. Snihirova et al., "Galvanic corrosion of Ti6Al4V -AA2024 joints in aircraft environment: Modelling and experimental validation", *Corrosion Science Volume 157*, 2019, Pages 70-78, doi: 10.1016/j.corsci.2019.04.036.

3. C. Kandekar et al., "A partitioned computational framework for damage evolution in stresscorrosion cracking utilizing phase-field", Proceedings in Applied Mathematics & Mechanics, 2023, doi: 10.1002/pamm.202200211.

FAIR assessment practices

Experiences from KonsortSWD and BERD@NFDI.

Janete Saldanha Bach¹ [<https://orcid.org/0000-0001-9011-5837>], Fidan Limani² [<https://orcid.org/0000-0002-5835-2784>],
Yudong Zhang¹ [<https://orcid.org/0009-0006-7108-475X>], Atif Latif² [<https://orcid.org/0000-0003-3085-3031>],
Brigitte Mathiak¹ [<https://orcid.org/0000-0003-1793-9615>], and Peter Mutschke¹ [<https://orcid.org/0000-0003-3517-8071>]

¹ GESIS - Leibniz-Institute for the Social Sciences, Germany

² ZBW – Leibniz Information Centre for Economics, Germany

Abstract. The poster presents FAIR assessment experiences in the context of the two NFDI consortia KonsortSWD and BERD@NFDI, employing the established Research Data Alliance's FAIR Data Maturity Model (RDA-FDMM) and the F-UJI Tool, an automated solution. RDA-FDMM, a manual technique, is more comprehensive, while the automated F-UJI tool effectively detects areas of improvement in metadata presentation that automated means can address. Our experiences highlight the need to examine both machine-readable as well as non-machine-readable elements and acknowledge automated tools' limitations, while valuing their insights. As the research ecosystem advances, metadata representation should be made increasingly machine-readable. We recommend a "FAIR by design" approach from the beginning to ensure alignment with FAIR principles in project outcomes. Continuous assessments during a project's lifetime promote ongoing research data infrastructure improvements within the NFDI consortia context, contributing to NFDI infrastructure innovation and optimization.

Keywords: FAIR principles, FAIR assessment, RDA FAIR Data Maturity Model, Automated FAIR assessment tool.

1. Introduction

The FAIR Data Principles [1] are widely applied to research data infrastructures. However, due to their interpretation scope, it is still challenging to assess the extent to which a data infrastructure addresses the FAIR principles. The Research Data Alliance has proposed a FAIR Data Maturity Model (RDA-FDMM) [2], based on which we share FAIR assessment experiences from KonsortSWD'sⁱ PID service [3] and BERD@NFDI'sⁱⁱ metadata schema. Moreover, we studied a FAIR automatic assessment using the F-UJI tool [4], which employs the RDA-FDMM and the FAIRsFAIR metrics [5] in a machine-readable fashion. We applied the F-UJI tool to GESIS Search in the context of KonsortSWD, motivated by the European landscape study [6] which also relies on F-UJI tool and led us to improve our metadata [7]. Our findings

ⁱ KonsortSWD (Consortium for the Social, Behavioural, Educational and Economic Sciences) is funded by the National Research Data Infrastructure (NFDI). KonsortSWD Homepage: <https://www.konsortswd.de/>

ⁱⁱ BERD@NFDI is an initiative to build a powerful platform for collecting, processing, analysing, and preserving Business, Economic, and Related Data. It is funded by the National Research Data Infrastructure (NFDI). BERD@NFDI Homepage: <https://www.berd-nfdi.de/>

represent a great opportunity to support convergence towards FAIR adoption and implementation at the NFDI level, while overcoming challenges related to different projects.

2. FAIRness Assessment Methodologies

FAIR assessments based on automatic tools [8] [9] [10], online self-assessment surveys [11], [12], [13], [14], manual [15] [16] [17] and hybrid methods [18] exist. The RDA-FDMM stands out as the most prominent manual method due to its completeness and comprehensive acknowledgement from the broader FAIR community. Most assessment tools rely partially on the RDA-FDMM model's indicators and measures. The RDA-FDMM defines 41 FAIR indicators, organized into three classes (*Essential*, *Important*, and *Useful*), and five levels (see Table 1).

Table 1. RDA-FDMM: indicators classes in five levels [19].

Classes	Indicators Quantity	Level 1	Level 2	Level 3	Level 4	Level 5
Essential	20	20	20	20	20	20
Important	14		7	14	14	14
Useful	7				3	7
Total according to the sum indicators		20	27	34	37	41

The RDA-FDMM measures indicators based on binary questions, by a "progress" evaluation, or by a hybrid mode. The KonsortSWD applied the binary method on each indicator, while BERD relied on a hybrid approach [2].

An example of an automated FAIR assessment is provided by the F-UJI tool, which, as a consequence, considers only indicators that can be assessed automatically, leading to a subset of just 16 [5] out of the 41 indicators proposed by RDA-FDMM.

3. FAIR Assessment using the RDA FAIR Data Maturity Model

3.1 KonsortSWD PID Service

In KonsortSWD we applied the RDA-FDMM to its PID service aiming to assign PIDs to data elements below study level (such as for survey variables). The PID service is based on the data registration agency da|ra [3], and the indicators were manually assessed at the PID service or at the da|ra level, using the pass-or-fail method. This approach is focused on determining how a resource under evaluation performs on meeting the indicators across the FAIR areas. In that sense, it gives a binary answer to each indicator. The results show that the PID service meets all the indicators classified as *essential* and most of the indicators from the classes *important* and *useful* (see Table 2).

Table 2. PID and da|ra service assessment results: level distribution.

Classes	Indicators Quantity	Level 1	Level 2	Level 3	Level 4	Level 5
Essential	20	20 / 20	20 / 20	20 / 20	20 / 20	20 / 20
Important	14		7 / 7	10 / 14	10 / 14	10 / 14
Useful	7				3 / 3	3 / 7
Achieved indicators		20/20	27/27	30/34	33/37	33/41
Scored		20	27	30	33	33
Results		100%	100%	88%	89%	80%

We fully comply with levels 1 and 2, achieve 88% compliance at level 3, 89% at level 4, and 80% at level 5. The failed indicators are concerned with automatic features, including references and/or qualified references to other data, and data accessed automatically (i.e., by a computer program).

3.2 BERD Metadata schema

In BERD@NFDI, we assessed the core part of the project's metadata schema, represented by DataCite Schema [20]. We wanted to assess the extent to which the elements of this schema can support the FAIR principles as identified in the RDA-FDMM, and not the metadata values of an (digital) object. Thus, our evaluation scope included only the FAIR principles that relate to the metadata aspects, resulting with 26 indicators in total, of which 14 *essential*, 9 *important*, and 3 *useful*. We applied the "0 - not applicable" score to the data-related indicators; Table 3 shows the indicators that pertain to both data and metadata.

Table 3. RDA-FDMM: Indicators/FAIR category for the data and metadata.

RDA Maturity Model Per Category	Indicators Per Category	Metadata Related	Data Related
F	7	5	2
A	12	6	6
I	12	7	5
R	10	8	2
Total	41	26	15

The assessment results show a relatively low FAIRness progress for the A and I categories of FAIR, whereas F and R perform better. The FAIR principles encompass data, metadata, and infrastructure [21] and the lack of any entity during evaluation normally affects the assessment score. Thus, since the data-related indicators are not being considered for this case, this disfavours the final assessment score.

4. Lessons learned from using an automated FAIR Assessment Tool

As an example of an automated FAIR assessment, we used the F-UJI tool to assess the GESIS Search as a relevant repository in the context of KonsortSWD. The automated assessment allowed us to identify actions to improve our metadata and metadata representation by automated means. Implementing the measures led to a noticeable enhancement of our research data FAIRness, and to a set of recommendations to improve FAIRness scores. The recommendations include:

- ensure that the landing page is machine-readable, avoiding JavaScript generated contents;
- define available metadata in JSON-LD, both on the landing page and in the used PID registration system, e.g., DataCite;
- provide links to the content resources (e.g., the PD article, CSV datasets, etc.) on the landing page. Linked content resources of long-term readability such as plain text are preferred;
- ensure metadata for linked data is correct and complete;
- use the standards suggested by F-UJI to complement free-form descriptions;
- keep your re3data record up to date and define an OAI-PMI endpoint for it.

It is essential to highlight that automatic tools can support FAIRness evaluation only partially. Although automation saves effort, not all components of the research data ecosystem are machine-readable. Some FAIR principles' aspects still require human mediation and interpretation [22]. On the other hand, using a tool like F-UJI is valuable in identifying weaknesses in metadata presentation that can be improved by automatic means, e.g., when the required metadata do not exist in machine-readable way, such as metadata generated by JavaScript.

We propose a "FAIR by Design" approach which, following Privacy by Design (PbD) [23] [24] where privacy measures are embedded directly into technology and business practices

from their inception. "FAIR by Design" aims to align research data infrastructures with FAIR principles through their entire lifecycle. Regular FAIR assessments and continuous improvements in FAIR scores should become an integral part of any data infrastructure development.

5. Conclusion and recommendations

The RDA-FDMM is a comprehensive standard for manual FAIR assessment broadly recognized by the FAIR community. The in-depth FAIR analysis using RDA-FDMM helped us better understand where our services stand with regards to FAIR, whereas the F-UJI tool gave us valuable hints on how to improve our metadata, despite the fact that automated tools always have limitations and technical challenges. Our experience with RDA-FDMM and the F-UJI tool highlights the importance of evaluating both machine-readable as well as non-machine-readable elements. Thus, we considered both cases in our study.

There are two major take-aways from our studies: (1) Apply both broader standards for manual FAIR assessment like RDA-FDMM, as well as automated tools like F-UJI to finally get a comprehensive picture regarding the FAIR compliance of research data infrastructures. (2) Adopt a "FAIR by design" approach early in product or service development to ensure that the FAIR principles are embedded in the development of research data infrastructures from the beginning, including regular FAIR assessments throughout the project lifetime to evaluate how the ongoing improvement of research data infrastructures affects the FAIR maturity score. This approach should be applied to NFDI as well, to finally innovate NFDI infrastructures.

Author contributions

The listed authors have prepared and written this extended abstract (role: Writing – original draft according to [CreDIT guidelines](#), Contributor Roles Taxonomy).

Competing interests

The authors declare that there are no competing interests.

Funding

KonsortSWD is funded by the German Research Foundation (DFG) within the framework of the NFDI – project number: 442494171.

BERD@NFDI is funded by the German Research Foundation (DFG) within the framework of the NFDI – project number: 460037581.

References

1. M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, no. 1, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
2. Research Data Alliance FAIR Data Maturity Model Working Group, "FAIR Data Maturity Model: specification and guidelines," 2020, doi: 10.15497/RDA00050.
3. C.-P. Klas, M. Zloch, J. S. Bach, E. Baran, and P. Mutschke, "KonsortSWD Measure 5.1: PID Service for variables report," Zenodo, Mar. 2022. doi: 10.5281/ZENODO.6397367.
4. A. Devaraju *et al.*, "From Conceptualization to Implementation: FAIR Assessment of Research Data Objects," *Data Science Journal*, vol. 20, p. 4, Feb. 2021, doi: 10.5334/dsj-2021-004.
5. A. Devaraju and P. Herterich, "D4.1 Draft Recommendations on Requirements for Fair Datasets in Certified Repositories," Feb. 2020, doi: 10.5281/ZENODO.3678715.
6. European Commission. Directorate General for Research and Innovation., Visionary Analytics., DANS., DCC., and EFIS., European Research Data Landscape: final report. LU: Publications Office, 2022. Accessed: Aug. 18, 2023. [Online]. Available: <https://data.europa.eu/doi/10.2777/3648>.
7. J. S. Bach, C.-P. Klas, B. Mathiak, Y. Zhang, and P. Mutschke. "FAIRness assessment: a comparison of the RDA model and the F-UJI Automated tool report," Zenodo, 2023, doi: 10.5281/zenodo.8308902.
8. M. D. Wilkinson *et al.*, "Evaluating FAIR maturity through a scalable, automated, community-governed framework," *Sci Data*, vol. 6, no. 1, p. 174, Sep. 2019, doi: 10.1038/s41597-019-0184-5.
9. T. Rosnet, V. Lefort, M.-D. Devignes, and A. Gaignard, "FAIR-Checker, a web tool to support the findability and reusability of digital life science resources," Jul. 2021, doi: 10.5281/ZENODO.5914307.
10. A. Ganske *et al.*, "ATMODAT Standard (v3.0)," 2021, doi: 10.35095/WDCC/ATMODAT_STAN_DARD_EN_V3_0.
11. Universidade Federal da Paraíba (PPGCI/MPGOA - UFPB), "FairDataBR: a tool for datasets evaluation". <https://wrco.ufpb.br/fair/>.
12. Australian Research Data Commons (ARDC), "ARDC FAIR data self-assessment tool," 2022, <https://ardc.edu.au/resource/fair-data-self-assessment-tool>.
13. Data Archiving and Networked Services, "SATIFYD: self-assessment tool to improve the fairness of your dataset," <https://satisfyd.dans.knaw.nl>.
14. Data Archiving and Networked Services, "Fairaware: your first step towards your FAIR data(set)," <https://fairaware.dans.knaw.nl/>.
15. WDS/RDA Assessment of Data Fitness for Use WG, "WDS/RDA Assessment of Data Fitness for Use WG Outputs and Recommendations", doi: 10.15497/RDA00034.
16. C. Bahim *et al.*, "The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments," *Data Science Journal*, vol. 19, p. 41, Oct. 2020, doi: 10.5334/dsj-2020-041.

17. J. Yu and S. Cox, "5-Star Data Rating Tool." CSIRO, 2017. doi: 10.4225/08/5A12348F8567B_
18. D. J. B. Clarke *et al.*, "FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources," *Cell Systems*, vol. 9, no. 5, pp. 417–421, Nov. 2019, doi: 10.1016/j.cels.2019.09.011_
19. J. Saldanha Bach, C.-P. Klas, and P. Mutschke, "Application of 'RDA FAIR Data Maturity Model' to assess the PID registration service in terms of FAIRness," Oct. 2022, doi: 10.5281/ZENODO.7409651_
20. DataCite Metadata Working Group, "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4," p. 82 pages, 2021, doi: 10.14454/3W3Z-SA82_
21. E. Schultes, Official GO FAIR Foundation icons for the Three-Point FAIRification Framework. Zenodo, 2021. doi: 10.5281/ZENODO.4678333.
22. A. Devaraju and R. Huber, "F-UJI : An Automated Assessment Tool for Improving the FAIRness of Research Data," Sep. 2020, doi: 10.5281/ZENODO.4068347_
23. A. Cavoukian, "Privacy by Design (PDF)", Information and Privacy Commissioner, https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf.
24. P. Hustinx, "Privacy by design: delivering the promises," *IDIS*, vol. 3, no. 2, pp. 253–255, Aug. 2010, doi: 10.1007/s12394-010-0061-z_

Introducing JuOSC

Monica Gonzalez-Marquez¹ and Ines Schmah¹ [<https://orcid.org/0009-0003-5878-3706>]

¹ Forschungszentrum Jülich

Abstract. Where does one go to find information on Open Science? At the moment, the literature is scattered across the various fora that comprise scientific discourse. An inquiry about any topic in Open Science, for example, how to manage a particular type of data in chemistry or what the current stance of the DFG or the NSF might be on data on photovoltaics or the relationship between Kuhn and Open Science, invariably require a mixture of asking colleagues or conducting random searches in academic indexing services, with varying degrees of success. Part of the problem is that so much of Open Science discourse has occurred in fora made possible by the advent of the internet. Between blog posts, twitter threads, talks and podcasts, there is a wealth of information that needs to be organized and archived to be most useful to the scientific community. There are efforts to organize some of this information in various online databases, and they are laudable, but to serve as long-term tools, they require 1. orchestration to gather and curate materials as a community effort, 2. long-term infrastructure to store copies of the materials, and 3. the expertise of librarians to manage such a collection.

Keywords: Special Collection, Open Science as a discipline

1. Our solution

We have begun to build the Jülich Open Science Collection (JuOSC) as the hub for all literature related to Open Science as a discipline. The collection will include traditional publication outputs (books and papers) as well as gray literature (blog posts, preprints, theses, etc.) and ephemera (twitter threads, podcasts, etc.). Because the collection is specifically for Open Science as a discipline, by definition all materials it contains will be freely available to everyone at no cost, as well as available on demand. As such, we are ensuring that all materials are either open access or available through a cc-by license. To ensure access on demand, copies of all materials will be stored on the JuSER infrastructure at Jülich. We will also construct mirrors at Zenodo and at the Internet Archive.

2. Implementation

JuOSC will be stored on the Jülich Shared Electronic Resources (JuSER), the publications database developed for Forschungszentrum Jülich. The Central Library maintains and develops the database based on the open source software [Invenio](#). The use of open source software ensures that modifications can be made by developers as needed by stakeholders, including the open science community, now and in the future. At present, JuSER provides a [list of document types](#). The complete workflow covers the collection of potential literature, the curation, the importing of metadata, archiving the files in JuSER and publishing to make the open science literature accessible worldwide to everyone. In addition, archiving workflows for all materials, including audio and visual records, are being developed in collaboration with experts from the [LZA-TREFF-KOELN](#), a network of people dedicated to addressing issues relating to long-term archiving. Because we are committed to creating a special collection that

serves the Open Science community *in collaboration with* the Open Science community, curation of materials is being done with the participation of the Open Science community. Specifically, this means that previously developed databases on Open Science literature will be integrated into JuOSC. Several such databases have already been identified, e.g. the [Open Research, Open Science, Open Scholarship](#) database developed by Erzsébet Tóth-Czifra and colleagues, and collaboration agreements are in development for others, i.e. Framework for Open and Reproducible Research Training ([FORRT](#)). We are also creating tools to facilitate input from the Open Science community such as suggestions for categories of literature and for particular items using a [Google Form](#) and hashtags for use on social media, i.e. #JuOSC. For all steps we aspire to an intensive exchange with the community in general and with any and all groups, experts and stakeholders who are willing to support our efforts. As to legal access to all materials we will archive, as stated above, all will be either open access or cc-by. Any materials that do not meet those requirements and for which an alternative is not available, i.e. a preprint instead of a paywalled paper, will be excluded. Our firm policy is that if an item is contained in our database, it will be freely available on demand, no exceptions.

3. Awareness

The purpose of JuOSC is not only to serve as information infrastructure, but to be actively used to build the open science community, both within Jülich and beyond. The same as with all disciplines, when scholars are first exposed to a new idea they begin with a literature search. The overarching goal of JuOSC is to facilitate education and exposure to new and established ideas and practices surrounding Open Science for all stakeholders. Our hope is that JuOSC will help the stakeholders at large to align their strategies and use limited time and resources more productively.

4. Conclusion

Our immediate aim is to develop a prototype by the end of the year, and to start using JuOSC to organize events and teach courses and workshops at the Forschungszentrum Jülich. Our approach will hopefully be useful to other institutions, and help strengthen the Open Science community by increasing awareness of and adoption of Open Science practices.

Competing interests

All authors declare there are no competing interests.

PIDs in the Natural Sciences

Thomas Schörner¹[\[https://orcid.org/0000-0002-7213-0352\]](https://orcid.org/0000-0002-7213-0352), Anton Barty¹, Markus Demleitner², Harry Enke³[\[https://orcid.org/0000-0002-2366-8316\]](https://orcid.org/0000-0002-2366-8316), Oliver Koepler⁴[\[https://orcid.org/0000-0003-3385-4232\]](https://orcid.org/0000-0003-3385-4232), Martin Köhler¹[\[https://orcid.org/0000-0003-0617-3319\]](https://orcid.org/0000-0003-0617-3319), Bridget Murphy⁵[\[https://orcid.org/0000-0002-1354-2381\]](https://orcid.org/0000-0002-1354-2381), Sonja Schimmler⁶[\[https://orcid.org/0000-0002-8786-7250\]](https://orcid.org/0000-0002-8786-7250) and Lisa-Marie Stein¹[\[https://orcid.org/0000-0001-7905-0462\]](https://orcid.org/0000-0001-7905-0462)

¹ Deutsches Elektronen-Synchrotron DESY, Germany

² Universität Heidelberg, Germany

³ Leibniz-Institut für Astrophysik Potsdam AIP, Germany

⁴ TIB Leibniz-Informationszentrum Technik und Naturwissenschaften Universitätsbibliothek, Germany

⁵ Christian-Albrechts-Universität zu Kiel, Germany

⁶ Fraunhofer-Institut für offene Kommunikationssysteme FOKUS, Germany

Abstract. We report on ongoing discussions and plans of NFDI consortia in physics and related natural sciences with respect to (persistent) identifiers.

Keywords: Persistent identifiers, Physical sciences, Natural sciences, FAIR principles, PID guidelines, Knowledge distribution

1. Introduction

Identifiers, and in particular persistent identifiers (PIDs), are a ubiquitous phenomenon, a desideratum, a challenge and a requirement in all branches of science, as e.g. reflected by NFDI working group of the NFDI section “Common Infrastructures” on the topic, and a topical proposal for the first call of the Base4NFDI initiative.

In the physical sciences – and with a little broader horizon: in the natural and technical sciences – such identifiers are used for a wide variety of entities, tangible and digital: data sets, publications, software, samples of certain materials or chemical compounds, hardware devices from individual chips to entire detectors, beamline instrumentations, ...

When choosing adequate identifiers to be used in the digital context, one has to consider several questions:

First and foremost, one has to clarify the purpose of the identifier, and its scope, or level of detail. If the identifier stands for a limited collection of data sets, there is almost always a lean procedure, available from the immediate scientific context, to provide order and identification. Lifting such a collection out of this context, for example through further processing steps, may make additional identifiers (and metadata) necessary. In German, there is the term “Erschließungstiefe” that summarises the required careful deliberation process on which information to carry on. This is particularly important if the identifier points to a real-world object (with its potentially infinite depth of detail).

The next question is associated with the purpose of the additional identifiers referred to above. Here, the FAIR principles can provide ample guidance. The answer should also contain information on whether there are already identifiers in use within the community / discipline in question that might easily serve the purpose.

The third question to consider is connected to the (intended or expected) persistence of both, the identifier itself and the object that is identified. For physical samples as objects, there is additionally the issue of the lifetime of the sample to be taken into account. For digital objects (e.g. datasets) their life-cycle must be taken into account and properties should match those of the digital identifier.

For the digital identifier itself it needs to be clarified if it should have the property to be persistent and resolvable. In all three cases the question of (institutional) commitment (e.g. DOI) must be raised.

The topic of (persistent) identification leads furthermore to the necessary awareness of defining the status of the data the scientist is dealing with. Because once it is classified which type of data we are addressing – ranging from primary to secondary, from raw experimental data to derived data sets – the question arises which outputs should be published and which should only be given a public status within the individual infrastructure. (Persistent) identifiers are a fundamental principle in the publication process that need to be further promoted. But also at this point it is necessary to create an awareness for their use, since type and use are also conditioned by the juxtaposition of *public* and *published*. In summary, for a community, the choice of identifiers also requires clarification of the criteria that determine the decision to publish an object.

2. Physical Sciences

In the physical sciences, a structured discussion on the matter of PIDs was started some time ago between players from various structures in the field:

- The NFDI consortia DAPHNE4NFDI and PUNCH4NFDI cover the disciplines of photon and neutron science, and of particle, astroparticle, hadron&nuclear physics and astrophysics. The issue of metadata and of identifiers plays a significant role in both their work programmes.
- Within the DIG-UM community self-organisation for the ErUM-Data action plan, eight sub-communities and their research committees are represented and work, among other things, on "Research Data" within the topic group with the same name: Astroparticle physics [1], elementary particle physics [2], accelerator physics [3], research with neutrons [4], research with synchrotron radiation [5], research with nuclear probes and ions [6], hadron and nuclear physics [7], Rat Deutscher Sternwarten [8].

There is, naturally, significant overlap between these structures.

3. Other Natural Sciences

In the context of natural sciences, NFDI4Chem and NFDI4Cat are bringing similar and additional aspects of PID applications to the table:

- Referencing data from Electronic Lab Notebooks, chemical reactions, collections of datasets derived from a unique sample;
- referencing analytical instruments and their configuration used to generate data.

4. Goals

The goals of the discussion are manifold:

- Knowledge distribution: The discussion between players from different branches of physics, and with representatives from neighbouring disciplines like chemistry, material science, etc. – is ideal for spreading knowledge on PIDs and of discussing common views and differences, thus contributing to a more informed use of identifiers.
- Use case collection: A collection of use cases is well suited to illustrate the breadth of PID applications and choices and can thus cross-fertilise expertise from various communities.
- Development of PID guidelines for the field, with recommendations for the large science system and the entire NFDI. To this end, the connection to the NFDI working group is maintained in the discussion.
- On the practical side, a list of guiding questions is being designed as a practical help for individual researchers that are confronted with the challenge of designing PIDs for their objects.

The topics mentioned above are being collected in a white paper, conceived as a living document that is constantly updated and maintained.

This contribution to the CoRDI conference will function as a progress report and summarise the status of the discussion on PIDs in the physical sciences and of the white paper, connecting this discussion to the wider NFDI scope.

Data availability statement

There are no data underlying this contribution, which is entirely based on information discussions among members of various NFDI consortia and NFDI sections.

Author contributions

TS and HE wrote the original draft, MD, OK, MK, LMS reviewed and edited it, AB, BM and SS reviewed and commented on it. All participated in the discussions that led to this submission.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the DAPHNE4NFDI, PUNCH4NFDI, NFDI4Cat and NFDI4Chem NFDI consortia.

The DAPHNE4NFDI project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the NFDI Funding Programme (grant no. 460248799).

The NFDI4Cat project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the NFDI Funding Programme (grant no. 441926934).

The NFDI4Chem project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the NFDI Funding Programme (grant no. 441958208).

The PUNCH4NFDI project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the NFDI Funding Programme (grant no. 460248186).

Acknowledgement

We kindly acknowledge the in-kind support of DESY, Universität Heidelberg, AIP, TIB, Fraunhofer FOKUS for the presented work.

References

1. "Komitee für Astroteilchenphysik KAT". <https://www.astroteilchenphysik.de/das-kat.html> (26 April 2023)
2. "Komitee für Elementarteilchenphysik KET". <https://www.ketweb.de> (26 April 2023)
3. "Komitee für Beschleunigerphysik KfB". <https://www.beschleunigerphysik.de/de/> (26 April 2023)
4. "Komitee Forschung mit Neutronen KFN". <https://www.sni-portal.de/de/nutzervertretungen/komitee-forschung-mit-neutronen> (26 April 2023)
5. "Komitee Forschung mit Synchrotronstrahlung KFS". <https://www.sni-portal.de/de/nutzervertretungen/komitee-forschung-mit-synchrotronstrahlung> (26 April 2023)
6. "Komitee Forschung mit nuklearen Sonden und Ionenstrahlen KFSI". <https://www.sni-portal.de/de/nutzervertretungen/komitee-forschung-mit-nuklearen-sonden-und-ionenstrahlen> (26 April 2023)
7. "Komitee für Hadronen- und Kernphysik KHuK". <https://khuk.uni-mainz.de> (26 April 2023)
8. "Rat deutscher Sternwarten RdS". <https://www.astronomische-gesellschaft.de/de/rat-deutscher-sternwarten> (26 April 2023)

Exchanging Research Data with SampleDB

Malte Deckers¹[\[https://orcid.org/0000-0001-9032-3062\]](https://orcid.org/0000-0001-9032-3062), and
Florian Rhiem¹[\[https://orcid.org/0000-0001-6461-9433\]](https://orcid.org/0000-0001-6461-9433)

¹Forschungszentrum Jülich GmbH, Germany

Abstract: Exchanging research data between systems is a vital task in research data management, especially in the context of collaborations. Therefore mapping data into standardized data descriptions and offering interfaces for efficient data exchange are required. The metadata database and electronic lab notebook SampleDB supports various ways of exporting data to other services, such as Dataverse, SciCat, and Jupyter-Hub, as well as file exports in open formats. The concept of SampleDB federations allows data exchange in loosely coupled federations of SampleDB instances.

Keywords: Data Exchange, Electronic Lab Notebook, Federation, Integration, Research Data Management

1 Motivation

Persisting research data in a findable, accessible, interoperable, and reusable way, therefore implementing the FAIR principles stated by Wilkinson et al. [1], is an essential requirement for efficient (re)use of data and reproducible work and thus an important part of the rules of good scientific practice.

Research data can be collected by and stored in various systems, from raw data outputs, log files, or measurement control software to entries in electronic lab notebooks (ELNs), data catalogs, and repositories designated for data publication. In particular, inter-institutional and interdisciplinary collaborations require an efficient exchange between these systems, which might be operated by different organizations and use various data descriptions, concepts, and software solutions. Differing data descriptions and the use of various metadata schemas, ontologies, and vocabularies can result in isolated data silos, complicating data discovery and access. As these systems may be tailored to specific requirements, they can not easily be replaced by a shared infrastructure. Therefore solutions for an efficient and FAIR data exchange, keeping access restrictions, and preventing inconsistent duplicates and missing information on different systems are required.

The open-source, web-based research metadata database and electronic lab notebook SampleDB [2] allows the definition of individual metadata schemas fitted to specific environments and use cases. These schemas allow the description of process-specific metadatasets using various datatypes and conditions, that have to be met by a new record and are validated on creation. These are completed by attached files, a location history, comments, and publications to track the entire lifecycle of a datum,

e. g. a sample. To be able to keep track of measurements performed at other facilities and to make the locally-defined metadata useful for users from other institutions, SampleDB requires such methods for exchanging data with instances at other institutions or with other ELNs or metadata catalogs.

2 Data Exports

Sharing data with partners or preparing it for publication requires different ways of data export. Depending on the requirements this might include simple file exports in open formats that can be easily interpreted by the receiving end, as well as making information directly accessible to other computer systems via APIs.

In SampleDB there are several file-based export methods, such as a PDF document containing the metadata or archives as `.zip`, `.tar.gz`, containing the full metadata, including location assignments, comments, linked publications, and files.

To exchange data with other ELNs, the export and import of files using the standardized `.eln` file format, as defined by The ELN Consortium in [3], can be used. These exports can include multiple related objects as well, for example, to describe a whole process flow.

Besides these file exports, there are also methods of direct data export to other software systems, e. g. for data publication or data exploration and analysis.

Objects can be directly exported to Dataverse [4] repositories, with the process-specific metadata exported from SampleDB represented using the EngMeta [5] *Process Metadata* metadata block. A researcher can decide which metadata and related files should be shared with the Dataverse and when a SampleDB record is exported, a draft dataset is created to be reviewed, extended, and published by the researcher.

Records can also be made available to the SciCat [6] data catalog, by mapping the data description used in SampleDB to the categories and metadata fields used in SciCat.

To facilitate data analysis using the metadata stored in a SampleDB instance, SampleDB also supports a JupyterHub [7] integration. JupyterHub is a web service that can run Jupyter notebooks for multiple users, which can then be used for data analysis and exploration. SampleDB can provide relevant metadata to a notebook template server, which can then combine it with predefined notebook templates to create ready-to-run notebooks on JupyterHub. This way experienced researchers, e.g. instrument scientists, can prepare templates that allow guest scientists to more easily analyze and explore their data.

SampleDB also provides an HTTP API that allows users to implement custom export programs for systems that neither support the export file formats nor are supported by SampleDB directly.

3 Federation

The concept of SampleDB federations has been introduced in [8] to allow sharing of information in loose associations of SampleDB instances of collaborating institutions and facilities while keeping unique identifiers and tracking records across institutional boundaries.

This is accomplished by assigning universally unique identifiers (UUID) to SampleDB instances, which – combined with the unique identifiers of the objects – create a unique identifier for every dataset in a federation. These unique identifiers allow for keeping valid references within the federation, even when not all referenced information is shared. Authentication between databases is accomplished by exchanging pair-wise tokens when setting up a collaboration in a federation so that the federation can grow organically as its institutions and their researchers collaborate.

When a record is released to a federated instance it is first processed, for example, to add license information or to prevent personal or confidential data from being shared. Therefore the shared data and schema might not be exact copies of the original data and are transferred into an export schema. On the receiving side, data is then checked for consistency and validity to be accepted or declined.

In [8] a protocol to update imported data and send back the changes to the origin is proposed as well. To prevent conflicting object versions within the federation, the instance that created a record can review changes before accepting and probably redistributing them.

The federation API and data exchange format could as well be used by other services to be integrated into a more heterogeneous federation of metadata catalogs or electronic lab notebooks. However, a less complex data ex- or import, like the methods described above, might be more suitable and practical in many cases.

Underlying and related material

- SampleDB source code: <https://github.com/sciapp/sampledb>, DOI: [10.5281/zenodo.4012175](https://doi.org/10.5281/zenodo.4012175)
- SampleDB documentation: <https://go.fzj.de/sampledb>

Competing Interests

The authors declare that they have no competing interests.

Funding

–

Acknowledgements

We thank Daniel Kaiser for his contributions, especially in code review.

References

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, “The fair guiding principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, 160018, Mar. 2016. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18). [Online]. Available: <https://doi.org/10.1038/sdata.2016.18>.
- [2] F. Rhiem, “SampleDB: A sample and measurement metadata database,” *Journal of Open Source Software*, vol. 6, no. 58, p. 2107, 2021. DOI: [10.21105/joss.02107](https://doi.org/10.21105/joss.02107). [Online]. Available: <https://doi.org/10.21105/joss.02107>.

- [3] The ELN Consortium. "TheELNFileFormat." (2022), [Online]. Available: <https://github.com/TheELNConsortium/TheELNFileFormat> (visited on 04/21/2023).
- [4] M. Crosas, "The dataverse network: An open-source application for sharing, discovering and preserving data," *D-Lib Magazine*, vol. Volume 17, 2011. DOI: [10.1045/january2011-crosas](https://doi.org/10.1045/january2011-crosas). [Online]. Available: <https://doi.org/10.1045/january2011-crosas>.
- [5] B. Schembera and D. Iglezakis, "EngMeta - Metadata for Computational Engineering," *International Journal of Metadata, Semantics and Ontologies*, vol. 14, pp. 26–38, DOI: [10.1504/IJMSO.2020.107792](https://doi.org/10.1504/IJMSO.2020.107792). [Online]. Available: <https://doi.org/10.1504/IJMSO.2020.107792>.
- [6] SciCat Project, *SciCat Project*. [Online]. Available: <https://scicatproject.github.io/> (visited on 04/26/2023).
- [7] Project Jupyter, *JupyterHub*. [Online]. Available: <https://jupyter.org/hub> (visited on 04/26/2023).
- [8] M. Deckers, "Entwicklung eines föderierten Datenbanksystems zur verteilten Verwaltung von Forschungsdaten," Master's Thesis, FH Aachen, 2021. [Online]. Available: <https://juser.fz-juelich.de/record/904981>.

The Data Steward Service Center (DSSC)

FAIRagro RDM-expertise hub

Nikolai Svoboda¹[\[https://orcid.org/0000-0003-3860-4400\]](https://orcid.org/0000-0003-3860-4400), Lucia Vedder²[\[https://orcid.org/0000-0002-8924-9800\]](https://orcid.org/0000-0002-8924-9800),
Franziska Böhm⁴, Markus Möller⁶[\[https://orcid.org/0000-0002-1918-7747\]](https://orcid.org/0000-0002-1918-7747), Elena Rey-Mazón⁵[\[https://orcid.org/0000-0003-4813-5927\]](https://orcid.org/0000-0003-4813-5927), Marcus Schmidt¹[\[https://orcid.org/0000-0002-5546-5521\]](https://orcid.org/0000-0002-5546-5521), Birte Lindstädt³[\[https://orcid.org/0000-0002-8251-1597\]](https://orcid.org/0000-0002-8251-1597), and Ulrike Stahl⁶[\[https://orcid.org/0000-0002-5659-910X\]](https://orcid.org/0000-0002-5659-910X)

¹ Leibniz Centre for Agricultural Landscape Research (ZALF) <https://ror.org/01ygyzs83>, Germany

² University of Bonn (UBN) <https://ror.org/041nas322>, Germany

³ ZB MED Information Centre Life Sciences <https://ror.org/0259fwx54>, Germany

⁴ FIZ Karlsruhe – Leibniz Institute for Information Infrastructure (FIZ) <https://ror.org/0387prb75>, Germany

⁵ Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) <https://ror.org/02skbsp27>, Germany

⁶ Julius Kühn Institute (JKI) - Federal Research Centre for Cultivated Plants <https://ror.org/022d5qt08>, Germany

Abstract. The Data Steward Service Center (DSSC) is the central institution within FAIRagro to develop data management tools based on the needs of the scientific community. The DSSC organizes the continuous exchange of RDM knowledge and experience with other institutions, channels user requests from the community, and transfers knowledge from the FAIRagro task areas to the FAIRagro data stewards. FAIRagro data stewards are experts in the field of RDM for agrosystems research supervising and will train data curators in our community. Data stewards have core competencies in research data management (e.g., cross-scale from genes, phenomics, management to region; sensitive data, remote sensing, time series, plant, soil and related FAIRagro data). Knowledge and expertise is pooled to provide the full range of expertise to the community in one place to foster the coalescence of the community. The DSSC is headed by a coordinator and will house five data stewards, who are active in the community e.g. train data curators, give legal support. In the course of the project, further institutional or project data stewards will be integrated and the pool of experts will be further expanded. The network to the other NFDI consortia is continuously growing.

Keywords: Knowledge transfer, Help desk, DSSC, Data Stewards, Community, Network, RDM

1. FAIRagro

FAIRagro is a community-driven RDM initiative and focuses on the well-organized field of agricultural systems domain integrating important disciplines and scales needed to develop sustainable crop production and agroecosystems. FAIRagro provides researchers with FAIR and quality-assured RDM to generate, publish, and access relevant data, innovative RDM services, and modern data driven science methods to support and advance agrosystems research. FAIRagro is very well networked within the agricultural sciences and open towards NFDI and beyond.

1.1 The Data Steward Service Center (DSSC)

The data stewards networked in the DSSC are well trained in the area of general research data management and have specialized skills to respond to the needs of agrosystem science. Specific RDM expertise is provided at DSSC by the FAIRagro partners: e.g. Guidance on long-term storage of plant breeding data, use of DMPs, implementation of Minimum Information About a Plant Phenotyping Experiment (MIAPPE) (partner: IPK Gatersleben); FAIR data publication, data repository, standardized metadata, DOI, long-term archiving, data acquisition (at ZALF, the FAIRagro coordinating institution); intellectual property rights in distributed information infrastructures, copyright, licensing, data protection law and IT security, legal advice (partner: FIZ Karlsruhe), curation of spatial data in spatial data infrastructures (SDI) considering quality aspects and provenance, enabling SDI interoperability (partner: JKI Quedlinburg), management of large amounts of heterogeneous data originating from field robots and drones (Farming 4.0), knowledge transfer, data visualization (partner: University of Bonn).

The work of data stewards is mainly to support scientists as authors in RDM and especially in using the FAIRagro infrastructure and publishing data in qualified, domain-specific repositories.

1.2 Connecting the DSSC to the world: Our Online-Help Desk on the FAIRagro Portal

FAIRagro offers first level support consisting of training and information materials available via the FAIRagro Portal. The data stewards work in the area of second level support, easily reached via the help desk by several means such as e-mail, an online form, an implemented semi-automated chat-application or by phone. A third level support consists of direct contact with the developers in FAIRagro for very specific demands. Aim is to offer services quickly, easily accessible, in person and with a high degree of availability.

In addition, data stewards are temporarily assigned as local supporters in the community according to the principle of "Book A Data Steward" based on their skills and the need in the six FAIRagro use cases to implement professional data management from the beginning and act as trainers for the community.

1.3 Insight into the agricultural science community

FAIRagro and the DSSC largely build on the needs of the agricultural science community. As early as 2020, a large-scale survey was conducted to identify RDM needs and wishes of the target group. The results have been published [2] and scientifically evaluated [3]. They are the foundation of actions implemented in the DSSC. The results of an additional survey in 2023 are comparable [1], e.g. the urgent desire for a low-threshold, competent personal support.

Affiliation with DFG subject groups [4] reflects the expected diversity of agrosystems sciences, with most participants in soil science (23%), plant breeding and crop production (18% each), and ecology (10%). Others were reported at 13%.

The future needs of the FAIRagro community point to a very broad range of data and specific support services and training opportunities for these. In particular, a need was expressed for omics, laboratory, remote sensing, and time series data. The types of data to work with vary from numeric data (31%), geographic data (18%), and text data (16%). It is important to note that a small portion will work with non-digital data and source code (7% each).

1.4 A tight Network within NFDI

FAIRagro is well connected within the NFDI and with the NFDI Directorate, and its partner institutions are members of the NFDI Association. The FAIRagro consortium supports the idea

of the Research Data Commons [5; 6], will network the infrastructures and services used and align them with this concept. FAIRagro is THE NFDI hub for agricultural research data and is networked with the broader agricultural science community. FAIRagro is involved in many NFDI consortia (NFDI4Earth, DataPLANT, NFDI4BioDiversity, NFDI4Health, NFDI4Objects, NFDI4DataScience, NFDI4Culture, NFDI4Memory, NFDI4CS, MaRDI, and NFDI4Chem), enabling direct information exchange towards the DSSC. FAIRagro has agreed with NFDI4Earth, NFDI4BioDiversity, DataPLANT, NFDI4Microbiota, NFDI4Health, NFDI4Objects, and NFDI4DataScience on cross-consortia activities and developments such as the data steward's service in research data management.

1.5 Outlook

Ultimately, the innovative structure of a DSSC aims to achieve a high level of RDM awareness in the agricultural-science community and beyond. With our Help Desk on the FAIRagro Portal, we will provide a low-threshold, user-friendly service hub where people can share and find data, learn about RDM through documents and in person and get support in all stages of the data life cycle. Personal support of the data stewards can be attained at any time for a large range of agrosystem-related topics from data-management plans to legal issues, big data, geodata, metadata and more. Our goal is to steadily develop greater expertise in specific RDM needs of agrosystem subfields and to create a network beyond our field and together with other RDM hubs which is easy to be navigated along for users of our data services.

Data availability statement

- FAIRagro survey 2021: <https://doi.org/10.5073/20211013-105447> (data)
- DaKA community survey 2023: <https://doi.org/10.20387/BONARES-STHV-TY43>

Author contributions

Nikolai Svoboda: conceptualization, writing – original draft, writing – review & editing

Lucia Vedder: review & editing

Franziska Böhm: review & editing

Markus Möller: review & editing

Elena Rey-Mazón: review & editing

Marcus Schmidt: conceptualization, writing – original draft, writing – review & editing

Birte Lindstädt: review & editing

Ulrike Stahl: funding acquisition, project administration, writing – review & editing

Competing interests

The authors declare that they have no competing interests.

Funding

NFDI consortia are funded by the Deutsche Forschungsgemeinschaft DFG: FAIRagro grant no. no. 501899475.

Acknowledgement

Xenia Specka and the coordination and management team of FAIRagro.

References

1. Svoboda, N., Stahl, U., Möller, M., Everwand, R., Bracke, J., Duttmann, R., Kuhwald, M., Wamhof, T., Bock, D., Tapken, H., Herrmann, L., & Stiene, S. (2023). 2023 survey of agricultural science community needs: data use, data competency, support. (Version 1.0) [Data set]. Leibniz Centre for Agricultural Landscape Research (ZALF). <https://doi.org/10.20387/BONARES-STHV-TY43>
2. Senft, M., Stahl, U., Svoboda, N., 2021. Dataset: survey about research data management in agricultural sciences in Germany. <https://doi.org/10.5073/20211013-105447>
3. Senft M, Stahl U, Svoboda N (2022) Research data management in agricultural sciences in Germany: We are not yet where we want to be. PLoS ONE 17(9): e0274677. <https://doi.org/10.1371/journal.pone.0274677>
4. DFG subject groups: https://www.dfg.de/dfg_profil/gremien/fachkollegien/liste/index.jsp?id=207
5. Bierwirth, M., Glöckner, F.O., Grimm, C., Schimmler, S., Boehm, F., Busse, C., Degkwitz, A., Koepler, O., Neuroth, H., 2020. Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung.10.5281/zenodo.3895209.
6. Glöckner et al. (2020) Glöckner, F.O., Diepenbroek, M., Felden, J., Güntsch, A., Stoye, J., Overmann, J., Wimmers, K., Kostadinov, I., Yahyapour, R., Müller, W., Scholz, U., Triebel, D., Frenzel, M., Gemeinholzer, B., Goesmann, A., König-Ries, B., Bonn, A., Seeger, B., 2020. NFDI4BioDiversity - A Consortium for the National Research Data Infrastructure (NFDI).10.5281/zenodo.3943645.

Building the Next Generation of Data Savvy Biomedical Researchers

Jens Dierkes¹[\[https://orcid.org/0000-0002-0121-9261\]](https://orcid.org/0000-0002-0121-9261), Birte Lindstädt²[\[https://orcid.org/0000-0002-8251-1597\]](https://orcid.org/0000-0002-8251-1597), Ulrich Sax³[\[https://orcid.org/0000-0002-8188-3495\]](https://orcid.org/0000-0002-8188-3495), Canan Hastik³[\[https://orcid.org/0000-0003-1729-4642\]](https://orcid.org/0000-0003-1729-4642), Julia Fürst², Tanja Hörner⁶[\[https://orcid.org/0000-0003-3280-6941\]](https://orcid.org/0000-0003-3280-6941), Sebastian Klammt⁷[\[https://orcid.org/0000-0001-7852-4769\]](https://orcid.org/0000-0001-7852-4769), Ines Per-rar⁸[\[https://orcid.org/0000-0002-2830-6322\]](https://orcid.org/0000-0002-2830-6322), Iris Pigeot^{9, 6}[\[https://orcid.org/0000-0001-7483-0726\]](https://orcid.org/0000-0001-7483-0726), Katja Restel¹, Carsten Oliver Schmidt¹⁰[\[https://orcid.org/0000-0001-5266-9396\]](https://orcid.org/0000-0001-5266-9396), Aliaksandra Shutsko²[\[https://orcid.org/0000-0002-7091-5084\]](https://orcid.org/0000-0002-7091-5084), Dagmar Waltemath¹⁰[\[https://orcid.org/0000-0002-5886-5563\]](https://orcid.org/0000-0002-5886-5563), Atinkut Zeleke¹⁰[\[https://orcid.org/0000-0001-7838-9050\]](https://orcid.org/0000-0001-7838-9050)

¹ University and City Library Cologne, University of Cologne, Germany

² ZB MED Information Centre Life Sciences, Germany

³ Technical University Darmstadt, Germany

⁴ University Medical Center Göttingen, Germany

⁶ University of Bremen, Germany

⁷ KKS-Netzwerk e.V., Germany

⁸ Institute for Nutrition and Food Sciences, Bonn University, Germany

⁹ Leibniz Institute for Prevention Research and Epidemiology – BIPS, Germany

¹⁰ Institute for Community Medicine, University Hospital Greifswald, Germany

Abstract. Modern research data management in biomedicine requires data literacy skills. The NFDI4Health consortium, the national research data infrastructure for personal health data, addresses this need with several different training concepts for clinical and epidemiological scientists. Both the institutional anchorage of the training and the thematic focus vary and can be assembled from a modular system as required. The aim is to enable multipliers to adapt and use the training modules through open educational resources. In addition, a competency profile for data stewards is under development to support the choice of required training in research institutions. The results will be fed into the NFDI “Education & Training” Section and will contribute to the Data Literacy Alliance in the future.

Keywords: Data Literacy, Training, Life Sciences, Health Informatics, Promoting RDM

Introduction

Researchers, but also allied professionals in the biomedical and health fields need the skills to manage research data in a planned way that ensures transparency, quality and compliance with regulatory requirements. To equip them with the appropriate understanding and the necessary skills in research data management (RDM) and data science, the National Research Data Infrastructure for Personal Health Data (NFDI4Health, [6]) offers a modular, reusable training programme that is tailored to different audiences based on the FAIR principles for data stewardship [12]. Offerings range from a university-based, certified courses with multiple tracks for Data Scientists and Data Stewards (DS), to single units on specific topics or NFDI4Health services.

The design and delivery of trainings is accompanied by a DS pilot programme that specifically explores the role of academic libraries working with discipline-specific research institutions as multipliers in delivering NFDI4Health services to users [10].

The concepts and materials are designed to be reusable and adaptable, and fit with plans to build a more sustainable FAIR data literacy (DL) training environment across all NFDI consortia with the NFDI "Training & Education" Section (EduTrain). NFDI4Health will contribute to the related Data Literacy Alliance (DALIA) infrastructure, which is responsible for the continuous offering, demand-oriented, methodical, and technical development of learning and teaching materials in the NFDI e.V.

Methods

NFDI4Health follows a multi-level approach, targeting different audiences within the clinical, epidemiological and public health sciences: (i) graduate students, (ii) researchers and (iii) professionals such as technicians, study nurses or librarians. This approach leads to three strands of action:

1. Cross-domain graduate training programme for doctoral students (Data Train)
2. Modular training offers based on dynamic learning paths concerning biomedical RDM basics and specific services/tools provided by NFDI4Health, accompanied by a knowledge base (NFDI4Health training handbook)
3. Concepts and materials developed published as open educational resources (OER), which are didactically enhanced to complement the RDM train-the-trainer programmes

Training needs are also informed by user research and immediate responses of participants via feedback forms. A key finding of the first survey was the need for basic RDM training and training on specific aspects of RDM. Initially, training will be developed along these two lines, with a focus on the basic biomedical RDM. The materials will be bundled into problem-based, contextualised and applied learning journeys.

To enable scaling up, the course materials and concepts will be made available as OER. These will allow for self-directed study. Furthermore, the provided materials can be adapted and reused in other contexts. As part of the DS pilot, a competency profile for DS in health will be developed to serve as a template for future DS education.

Results

Cross-domain graduate training: Data Train

The U Bremen Research Alliance [11], with the support of the Federal State of Bremen, has established the cross-institutional and cross-disciplinary training program "Data Train" [2, 5] for doctoral researchers from member institutions. Data Train pursues the mission of strengthening basic competencies in DL, RDM, and data science, while offering doctoral researchers a platform to build an interdisciplinary and interinstitutional network. The nine NFDI consortia represented in Bremen participate in the development and operation of the training courses. For example, members of NFDI4Health delivered a couple of data stories. In return, courses are opened for the NFDI-communities whenever possible.

Since 2021, the programme has gone through two passes with a total of 40 lecturers, 14 lectures, 17 workshops and data stories and 222 doctoral researchers participating. Since the programme was offered virtually, more than 2,600 participations were registered. Data Train will now be offered annually. It covers the entire data value chain and makes an important contribution to DL training for science as well as for the private sector [3].

Modular training offers

The NFDI4Health team has designed several RDM modules for researchers. These are regularly offered at conferences and in cooperation with individual research institutes (> 15 events). The courses cover RDM basics, biomedical aspects of RDM, and NFDI4Health-specific services.

The NFDI4Health digital services are being developed against a set of six representative use cases [7, 8]. Modules are designed, tested, and consolidated by DS of the pilot in close cooperation with the specialist researchers from the use cases. NFDI4Health will create a network of DS with different and complementary areas of expertise.

OER Material

A handbook detailing the above developments in training, the curricula and the main lessons learnt is in preparation.

All the underlying materials are designed as OER and will be published under a CC-BY licence, initially in the Repository of Life Science [9] and Zenodo.

Outlook

To enable DS or multipliers to adapt and implement the trainings and to train new DS, NFDI4Health will publish a DS concept alongside the training handbook, based on the experiences of the DS pilot. These also feed into the work of EduTrain, which aims to sustainably improve DL in science across the consortia [4].

Once DALIA, a knowledge base for the use and provision of FAIR data [1], is up and running, the NFDI4Health OERs can be integrated into it. NFDI4health, DALIA, and the Section "Ethical, Legal, and Social Aspects" (ELSA) are connected via a privacy working package, which will enable to sort and offer OER along distinct DL careers with the future option of working towards certificated or other qualification artefacts (e.g. micro-credentials). NFDI4Health will also be part of DALIA's network of curators and experts.

Author contributions

JD: conceptualisation, supervision, project administration, writing – review & editing

BL: conceptualisation, supervision, writing – review & editing

US: conceptualisation, writing – review & editing

CH, JF, TH, SK, IP, IPe, KR, COS, AS, DW, AZ: writing – review & editing

Competing interests

The authors declare that they have no competing interests.

Funding

We acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG): NFDI4Health no. 442326535; the Federal Ministry for Science and Education (BMBF) and EU's Reconstruction and Resilience Facility: DALIA: no. 16DWWQP07A; and support by the U Bremen Research Alliance and the Federal State of Bremen.

Acknowledgement

We thank the whole NFDI4Health consortium for their valuable input.

References

1. DALIA: Knowledge-Base für „FAIR data usage and supply“ als Knowledge-Graph. https://www.fst.tu-darmstadt.de/forschung_fst/zusammenarbeit_in_der_forschung/dalia/dalia_ueberblick.de.jsp (26.4.2023)
2. Data Train - Training in Research Data Management and Data Science. <https://www.bremen-research.de/data-train/> (26.4.2023)
3. Garbuglia, Federica, Saenen, Bregt, Gaillard, Vinciane, and Engelhardt, Claudia. (2021). D7.5 Good Practices in FAIR Competence Education (1.2). Zenodo. <https://doi.org/10.5281/zenodo.6657165>
4. Herres-Pawlis, Sonja, Pelz, Norbert Kockmann, Roger Gläser, Manuela Richter, Johannes Liermann, Jochen Ortmeyer, et al. "Sektionskonzept Training & Education zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.v.". Zenodo, April 21, 2022. <https://doi.org/10.5281/zenodo.6475541>; <https://www.nfdi.de/section-edutrain/?lang=en>
5. Hörner, Tanja, Frank Oliver Glöckner, Rolf Drechsler, und Iris Pigeot. 2021. „Disziplinübergreifendes Modell Zur Ausbildung Von Forschungsdatenmanagement Und Data Science Kompetenzen: ‚Data Train – Training in Research Data Management and Data Science““. *Bausteine Forschungsdatenmanagement*, Nr. 3 (Dezember). German:56-69. <https://doi.org/10.17192/bfdm.2021.3.8343>.
6. NFDI4Health. <https://www.nfdi4health.de/en/> (26.4.2023)
7. NFDI4Health Task Areas. <https://www.nfdi4health.de/en/about-us/task-areas.html> (26.4.2023)
8. NFDI4Health TA5 „Use Cases“. <https://www.nfdi4health.de/en/about-us/task-areas/ta5-use-cases.html> (26.4.2023)
9. Publisso - Repository for Life Sciences. <https://www.publisso.de/en/publishing/repositories/repository-for-life-sciences> (26.4.2023)
10. Shutsko, Aliaksandra, and Birte Lindstädt. 'Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten – NFDI4Health: Pilotprojekt zu Bibliotheken und Forschungsdatenkompetenzzentren als Multiplikatoren („Data Steward“)'. *GMS Medizin - Bibliothek - Information* 20, no. 3 (22 December 2020): Doc27. <https://doi.org/10.3205/mbi000484>.
11. The University of Bremen Research Alliance. <https://www.uni-bremen.de/research-alliance> (26.4.2023)
12. Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3 (15 March 2016): 160018. <https://doi.org/10.1038/sdata.2016.18>.

Spreading the Love for Mathematical Research Data

Tabea Bacher ¹[\[https://orcid.org/0000-0002-4480-5527\]](https://orcid.org/0000-0002-4480-5527), Christiane Görgen
²[\[https://orcid.org/0000-0002-6476-956X\]](https://orcid.org/0000-0002-6476-956X), Tabea Krause ²[\[https://orcid.org/0000-0001-7275-5830\]](https://orcid.org/0000-0001-7275-5830), Andreas Matt
³[\[https://orcid.org/0000-0002-6371-5121\]](https://orcid.org/0000-0002-6371-5121), Daniel Ramos ³[\[https://orcid.org/0009-0008-5707-0911\]](https://orcid.org/0009-0008-5707-0911), and
Bianca Violet ³[\[https://orcid.org/0009-0004-3463-9205\]](https://orcid.org/0009-0004-3463-9205)

¹Max Planck Institute for Mathematics in the Sciences (MPI MIS), Germany

²Leipzig University, Germany

³IMAGINARY gGmbH, Germany

Abstract: The Mathematical Research Data Initiative (MaRDI) is the NFDI consortium for the mathematics community. We outline some of the challenges we face in spreading a culture of mathematical research data to a large community and starting a cultural change. We highlight our approach to tackling these challenges and present some successful activities: a colorful newsletter, personal interviews, an entertaining rabbit, FAIR chocolate, and interactive formats.

Keywords: Research Data Management, Mathematics, Dissemination

1 The MaRDI project

The Mathematical Research Data Initiative (MaRDI) [1], [2] is the NFDI consortium for the mathematics community. Modern research in mathematics relies heavily on research data. Many areas of mathematics use not only pen, paper, and large libraries of books and articles but also software for running complex computations, modular libraries to build models, large experimental datasets to run statistical analysis or machine learning techniques, catalogs, and classifications of mathematical objects, etc.

MaRDI aims to furnish the necessary tools for efficient research data management in mathematics, but also to educate the mathematical community on the importance of and the benefits of following good practices of data management, and ultimately helping mathematicians and researchers to do their work easier and better. However, MaRDI is a complex project, it pushes some users out of their comfort zone, and for many researchers reflecting on tools to make your work more efficient seems like arid and bureaucratic duties.

2 Challenges and our approach

For some mathematicians, the modern data paradigm has entered stealthily and unnoticed. Many of them claim that “they use no data”, although they get surprised by realizing how many items they use can be considered “data”. Every theorem, every

formula, and every result is potentially a piece of research data. Even your own identity is data that must be curated.

To enable a cultural change is a long-term and tedious process. Our approach is to be very direct and professional about communicating the benefits of research data management while at the same time being very honest and inclusive. We involve the community - also on an international level - from the very beginning and establish communication channels that are colorful (in design), personal (in telling stories) and interactive (in its tools), and, wherever possible, entertaining. The idea is to change the dry image of adding the extra bureaucracy burden to researchers and make data care a useful and joyful necessity. We believe that spreading our own enthusiasm and motivation is key to reaching a tipping point for cultural change. In the next chapter, we give an overview of our most successful activities.

3 Activities spreading MaRDI

3.1 Make it colorful

We started "Math Data Quarterly" [3], a periodical newsletter to reach the mathematical community at large. We devoted four issues of the newsletter to the four FAIR principles (Findability, Accessibility, Interoperability, and Reusability), in connection with mathematics. We describe how mathematical research data is Found today, which requires persistent identification (e.g. DOI, ORCID), comprehensive catalogs (zbMATH, MathSciNet), repositories (e.g. Zenodo), and search engines that can explore and retrieve mathematical knowledge, not only articles and books. We discuss how mathematical data is Accessed, which includes publication models and licenses, but also access protocols and common interfaces to retrieve. We discuss practical problems of Interoperability between different platforms and propose solutions that MaRDI is implementing. We advocate for good practices of documenting data and metadata in order to make it Reusable, to make results verifiable, and to make the community work together. Our newsletter is prepared for an international readership, not limited to German researchers working in data management. We set our goal to become a resource for current trends in the field. The newsletter offers surveys, regular sections, videos, and colorful illustrations, see Figure 1.

3.2 Make it personal

Part of the newsletter are the "Data Dates" video interviews, see Figure 2. In an informal setting, researchers are invited to discuss personal experiences related to research data. We have diverse interview partners, from young researchers to Fields medal winners. Videos are recorded online and edited to be 5-8 minutes long, adding subtitles for accessibility.

Who are the people behind MaRDI, and what motivates them? We introduce you to the people who shape MaRDI with their expertise and vision. Every two weeks, an interview is published in the "Making MaRDI" series available on Twitter [4] and our website [5], see Figure 3.

3.3 Make it entertaining

We started the "MaRDI Movies" series of entertaining and informative videos. The first episode is called 'Mardy, the happy math rabbit' [6], see Figure 4. Follow Mardy



Figure 1. Illustrations by Constanza Rojas-Molina, licensed under CC BY-NC-SA 4.0.



Figure 2. Screenshots of our Data Date video series (from left to right: Christiane Görden (host), Johan Comelin, Ulrike Meyer Yang, Cedric Villani, and Elisabeth Bergherr).



Figure 3. Two MaRDI Makers.

through the pitfalls of reproducing software results: An introduction to software review in mathematics by Jeroen Hanselmann.

In 2022, we engaged with the public in a mini-symposium at the annual meeting of the German Mathematical Society [7] and had a booth with a large interactive media installation. A highlight was FAIR chocolate offered to anyone interested in chatting for a few minutes about math research data, see Figure 4. Similarly, we organized a ‘Pizza and Data’ event, where students talked about their experiences with research data while enjoying their slices of delicious pizza.



Figure 4. Left: Mardy, the rabbit looking at a proof of the Riemann hypothesis. Right: MaRDI at a conference booth, offering “FAIR”-trade chocolate.

3.4 Make it interactive

“Infrastructure for Mathematics” is a one-semester university course at University Leipzig on research data management aimed at mathematicians. The course is prepared and held by MaRDI team member Christiane Görden, the idea is based on [8]. The lectures are highly interactive and flexible. They welcome math undergraduates, graduate students, and early postdocs. Students can integrate the class into their final diploma examination. This is the first course of its kind and is a prototype for future MaRDI short courses.

“Love Data Week” [9] is an annual international celebration of data during the week of Valentine’s Day. In 2023, we created an interactive website [10] that allows you to play around with various mathematical objects and learn interesting facts about their file formats.

Next to the Barcamp format, we also organize reproducibility exercises: participants choose one publication that contains research data, especially software, and try to reproduce it. These sessions have led to lively discussions and a new perspective of the participants on the publication of their own research data.

References

- [1] The MaRDI consortium, *MaRDI: Mathematical Research Data Initiative Proposal*, May 2022. DOI: [10.5281/zenodo.6552436](https://doi.org/10.5281/zenodo.6552436).
- [2] The MaRDI consortium. “Mardi website.” (2023), [Online]. Available: <https://www.mardi4nfdi.de/> (visited on 04/25/2023).
- [3] The MaRDI consortium. “Mardi newsletter.” (2023), [Online]. Available: <https://www.mardi4nfdi.de/community/newsletter> (visited on 04/25/2023).
- [4] The MaRDI consortium. “Mardi twitter.” (2023), [Online]. Available: <https://twitter.com/mardi4nfdi> (visited on 04/25/2023).
- [5] The MaRDI consortium. “Making mardi.” (2023), [Online]. Available: <https://www.mardi4nfdi.de/community/making-mardi> (visited on 04/25/2023).
- [6] The MaRDI consortium. “Mardi movies: ‘mardy, the happy math rabbit’.” (2023), [Online]. Available: <https://vimeo.com/781454778/c131cd0a1a> (visited on 04/25/2023).

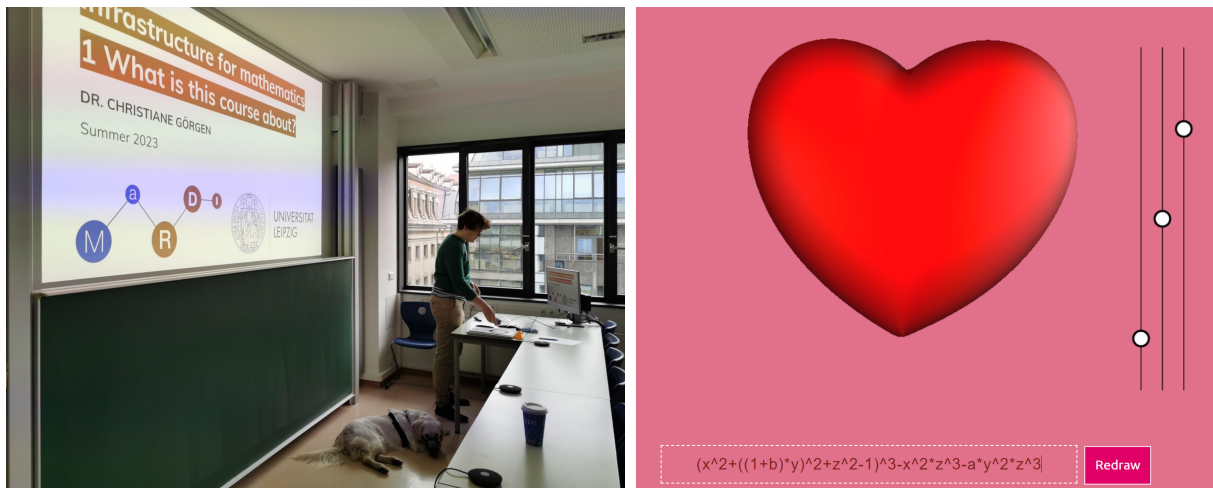


Figure 5. Left: A relaxed course on RDM. Right: An algebraic surface for Love Data Week.

- [7] Deutsche Mathematiker-Vereinigung (DMV). "Dmv annual meeting 2022." (2023), [Online]. Available: <https://www.mi.fu-berlin.de/dmv2022/program/minisymposia/index.html> (visited on 04/25/2023).
- [8] C. Wiljes and P. Cimiano, "Teaching research data management for students," *Data Science Journal*, Aug. 2019. DOI: [10.5334/dsj-2019-038](https://doi.org/10.5334/dsj-2019-038).
- [9] Inter-university Consortium for Political and Social Research (ICPSR). "Love data week." (2023), [Online]. Available: <https://www.icpsr.umich.edu/web/about/cms/3799> (visited on 04/25/2023).
- [10] The MaRDI consortium. "Mardi for love data week: 'a long and lasting love file format'." (2023), [Online]. Available: <https://www.mardi4nfdi.de/community/love-data-week> (visited on 04/25/2023).

DALIA FAIR Open Educational Federation

Aggregation, Harmonisation, Curation, and Quality Assurance with DALIA

Canan Hastik¹[\[https://orcid.org/0000-0003-1729-4642\]](https://orcid.org/0000-0003-1729-4642), Gábor Kismihók²[\[https://orcid.org/0000-0003-3758-5455\]](https://orcid.org/0000-0003-3758-5455), Frank Lange³[\[https://orcid.org/0000-0002-9346-6031\]](https://orcid.org/0000-0002-9346-6031), and Petra C. Steiner¹[\[https://orcid.org/0000-0001-8997-2620\]](https://orcid.org/0000-0001-8997-2620)

¹ Technical University of Darmstadt, Germany

² TIB - Leibniz Information Centre for Science and Technology, Germany

³ RWTH Aachen University, Germany

Abstract. As a funding measure of the German Federal Ministry of Education and Research and from the EU's Development and Resilience Facility, the project "DALIA - Knowledge-Base for FAIR data usage and supply" not only creates technical bridges between NFDI consortia but also an evolving approach that contributes to transdisciplinary community building in the long term. By harmonising, networking and making visible teaching and learning materials for the development and promotion of data literacy, the project makes an important contribution to education, training and cultural change.

Keywords: Training and Education, Knowledge Graph, Research Data Management, Harmonising RDM

1. Project Introduction and Objective

The DALIA project will develop a central entry point for a federated knowledge base, comprising teaching and learning materials to build and strengthen data literacy and knowledge in related FAIR [5] data use (data science) and supply (FAIR research data management). Thus, based on the DALIA Knowledge Graph, which is currently under development, the DALIA Knowledge-Base will make the heterogeneous, distributed and subject-specific teaching and learning materials of the NFDI e.V. [4], the Research Data Management (RDM) and Open Educational Resource (OER) communities visible, accessible, and networked in a sustainable way. By implementing the DALIA Knowledge-Base as Linked Open Data (LOD), not only the semantic interoperability of the learning content will be established but also the curation of teaching and learning materials. The DALIA Knowledge-Base will ensure optimal findability and best possible access for users from a wide range of disciplines, career and competence levels. It will also foster the (re)usability of the educational materials by means of a monitoring and personal learning recommendation system.

2. Challenges

As a federated infrastructure for FAIR data use and supply, DALIA addresses content providers and curators for educational purposes with learning content repositories such as NFDI4Ing's Education Repository (<https://git.rwth-aachen.de/nfdi4ing/education>), NFDI4Chem's Knowledge Base (https://knowledgebase.nfdi4chem.de/knowledge_base/),

DARIAH-Campus (<https://campus.dariah.eu/>), curated group repositories on Zenodo, e.g. Research data management (RDM) open training materials [2] or future data competence centers using their own community-specific taxonomies, for instance the Taxonomy of Digital Research Activities in the Humanities (TaDiRAH, <https://tadirah.info>). Similarly, to the European Open Science Cloud (EOSC, <https://eosc-portal.eu/>), DALIA wants to generate added value for users by networking these heterogeneous and distributed resources.

Challenges regarding the use of standards for material indexing, or the use of vocabularies for curation and search, are systematically tackled and harmonised across disciplines. In this context, characteristic questions are: How do we use vocabulary to harmonise resources for discovery and recommendation? What is the role of controlled vocabularies? Are keywords related to a discipline or resource? How is curation best organized? What is the common level of control?

The user-oriented findability, curation and quality assessment of teaching and learning materials in DALIA are the main cornerstones of the federated infrastructure. DALIA reuses existing open technologies, pipelines (e.g. data aggregator workflow in Europeana [1]), and assessment criteria (e.g. RDM learning goal matrix [3]) from FDM initiatives, which are working closely with NFDI e.V. and as well as communities from different RDM areas. These and the associated FAIR compliance are considered as important prerequisites for a trustworthy DALIA methodology, leading to the provision of a generic, but at the same time discipline-oriented educational service. In DALIA, the outcome of the harmonisation of metadata, including controlled vocabularies, will be represented in the DALIA Knowledge-Graph. A particular focus is to identify the scope and level of detail of controlled vocabularies and other metadata standards to be used to describe teaching and learning materials. Projects such as TaDiRAH and standards developed in focus group workshops and working groups such as the NFDI sections (Meta-)Data, Terminologies, Provenance, and Training & Education will be taken into account. This will yield networks of concepts which interlink to concepts from other resources. Controlled vocabularies, taxonomies and other ontologies, even if imperfect, are useful due to their semantic relations and expressiveness compared to other kinds of knowledge representation. They can be extended through language models and point to relevant similar concepts from other resources.

Interlinking learning content through DALIA is organized by content providers, who are responsible for the registration, quality assurance, and maintenance of learning content. Content curation is based on a small obligatory (closed) set of terms such as title, author, license, competency level, learning goal, topic/subject, link of the resource, and additional (open) keyword field to determine the precise context of a given learning resource.

3. Solution and Outreach

Thus, a common framework for the visibility of teaching and learning materials, their quality assessment and (re)usability is created. The aggregated content is curated in an RDF triplestore and is provided as Linked Open Data via a public SPARQL endpoint. In DALIA not only an API for an aggregated federated vocabulary service will be developed, but also a predefined set of common vocabularies for external repositories will be defined and a vocabulary charter with guidelines for DALIA users will be developed. This is intended to convey best practices for different curators, different metadata, and different vocabularies, and to encourage multipliers to assess quality.

Establishing interoperability is not a purely technical challenge. The harmonisation of the DALIA technology stack goes hand in hand with the development of framework conditions and didactic concepts for the quality-assured curation of practically usable teaching and learning materials for self-learning. To ensure the usability of DALIA in the long term, we will not only develop a train-the-trainer concept for learning content curation, but we will also collect user feedback in focus group workshops and via user monitoring continuously. The Data Literacy

Alliance behind DALIA ensures that the learning and teaching materials in the NFDI e.V. are continuously offered via the DALIA Knowledge-Base, and that they are further developed in terms of methodology and technology in line with requirements.

Data availability statement

This submission is not based on data.

Author contributions

CH - conceptualisation, writing original draft, review, and editing

GK – reviewing, and editing

FL – reviewing, and editing

PS - writing, reviewing, and editing

Competing interests

The authors declare that they have no competing interests.

Funding

This project with the federal label 16DWWQP07A is funded by the German Federal Ministry of Education and Research (BMBF) and by the EU's Reconstruction and Resilience Facility.

Acknowledgement

We would like to thank our DALIA project partners and related NFDI consortia for their valuable input.

References

1. Nuno Freire, Enno Meijers, Sjors de Valk, Julien A. Raemy, and Antoine Isaac, "Metadata Aggregation via Linked Data: Results of the Europeana Common Culture Project," in *Metadata and Semantic Research. 14th International Conference, MTSR 2020, Madrid, Spain, December 2–4, 2020, Revised Selected Papers* (Communications in Computer and Information Science 1355), Emmanouel Garoufallou, María-Antonia Ovalle-Perandones, Eds., Cham, Germany: Springer, 2021, pp. 383–394, doi: https://doi.org/10.1007/978-3-030-71903-6_35.
2. Laura Molloy. "Research data management (RDM) open training materials." <https://zenodo.org/communities/dcc-rdm-training-materials/?page=1&size=20> (26.04.2023).
3. Britta Petersen *et al.*, "Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards (Version 2)," 2023, Zenodo. doi: <https://doi.org/10.5281/zenodo.8010617> (29.08.2023).
4. York Sure-Vetter, Eva Lübke, Sophie Kraft, Hendrik Seitz-Moskaliuk, "Nationale Forschungsdateninfrastruktur (NFDI) e. V. - Satzungsvorstellung," 2021, doi: <https://doi.org/10.5281/ZENODO.5735196>.
5. Mark D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," in *Scientific Data*, vol. 3, Mar. 2016, Art no. 160018, doi: <https://doi.org/10.1038/sdata.2016.18>.

FDO to Structure the Domain of Knowledge

Peter Wittenburg¹[\[https://orcid.org/0000-0003-3538-0106\]](https://orcid.org/0000-0003-3538-0106) and Dimitris Koureas²

¹ FDO Forum, International

² Naturalis, Netherlands

Abstract. The Globally Integrated Dataspace will evolve and result in a domain full of digital artefacts and relations which in its size and complexity is unprecedented. Traditional methods of structuring the domain such as file structures are not sufficient any longer. On top of a layer of data centres and federations which need to implement mechanisms establishing trust, there will be layer of Virtual Collections serving different purposes. These VCs to a large extent need to be FAIR and persistent as well, since they include important information. We claim that the concept of FAIR Digital Objects is an excellent framework to build and manage these VC. Currently, there are a few implementations of the FDO concept.

Keywords: FAIR Principles, FAIR Digital Objects, Data Management

1. Introduction

The increasing volumes of data and their exponentially increasing interdependencies of different types create a complexity which is challenging for management and reuse in the evolving Globally Integrated Dataspace (GIDS) [1]. The widely agreed goal is that all different types of digital artefacts (data, metadata, software, semantic assertions, configurations, etc.) in this GIDS should be FAIR [2], and we also assume that we have mechanisms in place that guarantee improving responsibility, accountability, and persistence (RAP) for establishing trust. But this will not be sufficient for users to easily navigate and operate in this almost endless space. We will need to include structure at various levels to master the complexity.

Creators of digital artefacts create “canonical structures”, which are mostly hierarchical collections created with specific views in mind to help navigation and management. A typical tree structure in experimental science looks like “*experimenter/experiment-name/experiment-day/measurements*”. In observational research the structures that are being created by the creators are much more varied such as “*project/language/fieldtrip-name/day/observation-type/observations*”. In all cases, the individual observations or measurements are the leaves in such a tree. Another structuring element is typically the repository that hosts and manages this collection. In an IT sense, this just adds another layer to the chosen trees “*repository/project/language/...*”.

Assuming FAIR data, we can expect that increasingly “rich metadata” will be available describing the different leaves in the tree, i.e., in principle no-tree structure would have to be provided, but queries with a certain profile would help to get a result list which is an unstructured bag of digital artefacts. For some operations, such a list might be sufficient, but it will not help for many others. What we see often is that researchers want to create virtual collections following their own views on top of the leaves and perhaps reusing certain structural elements from canonical structures. With reusing digital artefacts from different disciplinary contexts for

different purposes, it is crucial to provide mechanisms to create, manage, exchange, and preserve such virtual collections. It should be noted though, that the leaves and even sub-structures in these virtual collections often will change over time – collections are living bodies.

2. Examples

In this section, we want to discuss a few examples of such recursively defined complex collections.

In the DOBES project a repository was created containing all material from about 250 researchers worldwide coming from highly different disciplines [3]. In addition to an agreed and well-defined metadata set, the creator teams wanted to define their own canonical hierarchical collection structures to simplify navigation and management, for example, to easily define rights. A simple search on specific types would not be sufficient. Also, the repository managers used the canonical tree for some operations, such as transforming all data items of type X in a certain sub-collection to a new type Y. But these canonical collection structures were not informative for other researchers who, for example, wanted to carry out an intonation analysis and comparison between languages. They wanted to create their own tree structure to facilitate the comparison and to document this structure to make it easily citable.

Another example of implying structure on a huge set of distributed data is the biodiversity digital specimen [4]. At many labs worldwide, different digital information about a specific physical object is being created, and it is related with many other objects across institutions due to a variety of classification schemes. In this example we have two canonical structures: (1) all information which is about a specific physical object, (2) different classification schemes to group the digital twins according to some criteria. In one case the leaves are the information sources about an object, in the other case the leaves are the digital twins. Also, in this case we can foresee that researchers want to carry out specific operations on virtual collections constructed according to their own criteria.

A last example shall briefly be indicated. Increasingly more people see the need to extract the major assertions in papers to create nano-publications, which are basically augmented RDF triples [5]. From insights about Medline, for example, it can be easily estimated that the number of such nano-publications will increase exponentially in the coming decade either by manual or automatic extractions. Again, we will need to create structures to be able to navigate or operate in such a huge space of semantic assertions. Different sets of criteria will be used to determine key concepts in such spaces as a start to form structures resulting in many different views on semantic spaces fit for specific purposes.

All these different virtual collections representing different views will have a high scientific value and many of them need to be preserved despite changes over time. They will be the basis of proper management and in addition, researchers and specialised brokers will put efforts in the creation of meaningful virtual collections that will have a value in themselves and will be reused, extended, changed etc.

3. Relevance of FDOs

FAIR Digital Objects (FDO) are atomic, self-containing units of information that persistently bundle all information needed for FAIRness [6]. They can be leaves but also collections due to their recursive definition which would mean that the body of each collection is the set of included elements, that the collection is assigned a PID and is associated with collection metadata. Therefore, FDOs are excellent mechanisms to organise this almost endless space of virtual structures, make them persistent, track their changes over time, and share them with others independent of the particular dataspace people are working in. FDOs are neutral with respect to the structuring of the body, i.e., it does not care how the different elements are described and referenced if the specifications will be machine actionable. Description standards such as RO Crate [7] could be used here. The type of the FDOs containing collections is

"collection" and a subtype could be "encoded_by_RO-Create" to enable machines to parse the structures.

Author contributions

Peter Wittenburg and Dimitris Koureas both contributed to the whole paper.

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

We would like to thank all the colleagues who contributed to the DOBES and DiSSCO work and the growing community pushing the FDO specifications and implementations.

References

1. P. Wittenburg, G. Strawn, "Shaping and standardising the Global Integrated Data Space", <https://sites.grenadine.co/sites/iot/en/iotweek-2022/schedule/8581/Shaping%20and%20standardising%20the%20Global%20Integrated%20Data%20Space>
The FAIR Guiding Principles for scientific data management and stewardship
2. M. Wilkinson, et al., "The FAIR Guiding Principles for scientific data management and stewardship", <https://www.nature.com/articles/sdata201618>
3. DOBES Archive, <https://dobes.mpi.nl/>
4. DiSSCO, "What is a digital specimen?", <https://dissco.tech/2020/03/31/what-is-a-digital-specimen/>
5. B. Mons, J. Veltrop, "Nano-Publication in the e-science era", <https://eur-ws.org/Vol-523/Mons.pdf>
6. G. Strawn, P. Wittenburg, "FDO Requirement Specifications", <https://zenodo.org/record/7781926#.ZEjWP87P3b0>
7. RO-Crate, "Research Object Crate", <https://www.researchobject.org/ro-crate/>

FAIRmat guide to writing data management plans

A practical guide for the condensed-matter physics and materials-science communities

Ahmed E. Mansour¹ [<https://orcid.org/0000-0002-3411-6808>], Lucia Rotheray¹ [<https://orcid.org/0009-0009-8969-1867>],
Kerstin Helbig² [<https://orcid.org/0000-0002-2775-6751>], Silvana Botti³ [<https://orcid.org/0000-0002-4920-2370>],
Heiko B. Weber⁴ [<https://orcid.org/0000-0002-6403-9022>], Martin Aeschlimann⁵ [<https://orcid.org/0000-0003-3413-5029>]
and Claudia Draxl¹ [<https://orcid.org/0000-0003-3523-6657>]

¹ Physics Department and IRIS Adlershof, Humboldt-Universität zu Berlin, Germany

² Computer and Media Service, Humboldt-Universität zu Berlin, Germany

³ Institut für Festkörpertheorie und -optik, Friedrich-Schiller-Universität Jena, Germany

⁴ Department of Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany

⁵ Department of Physics and Research Center OPTIMAS, Technische Universität Kaiserslautern, Germany

Abstract: Research data management is becoming an increasingly important topic due to the growing amounts of types, formats, and volumes of data produced by scientific research. In addition, a growing demand to make data accessible and comprehensible requires standardizing, managing, and planning the data life-cycle. For this reason, many funding agencies now require a data management plan (DMP) as part of submitted research proposals. While some of them and other scientific bodies offer DMP templates, there is no one-size-fits-all solution, due to the heterogeneity of data generated by different scientific disciplines. Here, we present as an example FAIRmat's effort in enhancing data literacy on the topic of DMP aiming to guide physicists and materials scientists to writing DMPs that comply with the requirements of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

Keywords: Data management plan, Data life-cycle, FAIR data, NOMAD, Materials Science

1. Introduction

Scientific research relies on the dissemination of research results. Traditionally, this process takes the form of conference contributions or publications. While such channels are currently accepted as the norm in the scientific community, challenges of reproducibility and accessibility arose along with concerns related to research integrity and compliance with ethical and legal requirements [1]. This has led to the advocacy of open access, the establishment of standards for reporting research methods and data availability statements to ensure reproducibility, and the enforcement of policies governing ethical and integrity issues [2–5]. As a result, there has been a paradigm shift in the research process from being linear – utilizing generated data towards a publication as the main outcome – to becoming a cyclic process – focusing on research data as key output as shown in Figure 1 [6,7]. Proper handling of research data requires a comprehensive management process, starting with planning prior to the project initiation, monitoring and controlling during the project, and defining the long-term fate of the data after the project closing. Such a management process is described by a data management plan (DMP).



Figure 1: Data life-cycle in research projects.

The German Research Foundation (DFG) now requires a detailed specification of the handling of research data to be included in research proposals [8]. This requirement is formulated as a catalogue of questions and is based on the criteria defined by the Science Europe association [9]. The DFG checklist is a high-level document that covers the diverse research disciplines [10]. It is crucial to adapt these requirements to each discipline, and it is up to the research communities to develop standards and best practices for preparing a domain-specific DMP.

2.The FAIRmat Guide

A core mission of FAIRmat is to support the scientific community to introduce and maintain high standards of reproducibility, research integrity, and compliance with ethical and legal requirements by adopting proper RDM practices based on the FAIR data principles [4]. A recent implementation of this mission is the [FAIRmat guide to writing a research data management plan](#) (see Figure 2), tailored for scientists in the fields of condensed-matter physics and materials science.



Figure 2: Front and back covers of FAIRmat's DMP guide.

The FAIRmat guide provides explanations of the elements of DMPs and practical tips for best practices in data management. It is prepared to meet the requirements for handling research data set by the DFG. The elements of the DMP described in the guide (as defined by the DFG requirements) are shown in Figure 3.

The preparation of this practical guide began with detailed discussions with established scientists in the field, covering both computational and experimental solid-state physics to identify the needs, currently available standards, and prospects for an effective DMP.

An effective DMP starts with a detailed description of what needs to be managed, i.e., the data generated or reused in the research project. This description forms the first section of the DMP, and guides all other sections that follow. Next, the plan should establish a comprehensive documentation for the project data through a structured organization and adopting a comprehensible naming convention for the data files. A key component is the use of standard and community-accepted metadata and ensuring the availability of rich metadata.

Management of the data storage, accessibly, and dissemination is described for both the short term (during the project) and the long term (after the project has ended). Our guide explains data repositories, their types and requirements, and refers the reader to resources for selecting a data repository suitable for subject of research and data type, such as the Science Europe criteria and the re3data directory [9,11].

Data description	<ul style="list-style-type: none"> Comprehensive description of data generated or reused in the project: data type, source, format and volume.
Documentation and data quality	<ul style="list-style-type: none"> A record of all the workflows for generating and structuring data, associating metadata, and how data quality will be ensured.
Storage and technical archiving	<ul style="list-style-type: none"> Methods and locations for secure and accessible storage of the data and detailed backup procedure.
Data exchange and long-term accessibility	<ul style="list-style-type: none"> Plan for the data fate after the project ends. It includes description of both preservation and dissemination.
Legal obligations and conditions	<ul style="list-style-type: none"> Identified legal issues relevant to the project data, such as: ownership, intellectual property, copyright and licensing laws.
Responsibilities and resources	<ul style="list-style-type: none"> List of tasks, assignments for project members, and costs relevant to data management

Figure 3: List of DMP sections and contents that meet the DFG requirements.

3. Conclusions

Our guide assists the research community in preparing an effective DMP to be submitted in accordance with the requirements of research proposals. This fills a gap in available resources on research data management for scientists. In addition to helping researchers to meet the minimum requirements for funding, our goal is to ensure the long-term success of research projects and improve the quality of the research process. Continuous improvement and regular updates are planned to ensure that the guide fulfills the latest standards and requirements of the DFG and other funding bodies.

Competing interests

The authors declare that they have no competing interests.

Funding

FAIRmat is a consortium of German National Research Data Infrastructure (NDFI) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project 460197019.

References

- [1] Deutsche Forschungsgemeinschaft (DFG), Guidelines for Safeguarding Good Research Practice. Code of Conduct, 2022.
- [2] European Commission, *The EU Open Science Policy*, https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en#documents.
- [3] M. Scheffler et al., *FAIR Data Enabling New Horizons for Materials Research*, *Nature* **604**, 635 (2022).
- [4] M. D. Wilkinson et al., *The FAIR Guiding Principles for Scientific Data Management and Stewardship*, *Sci Data* **3**, 160018 (2016).
- [5] L. M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lüders, M. Oliveira, and M. Scheffler, *Towards Efficient Data Exchange and Sharing for Big-Data Driven Materials Science: Metadata and Data Formats*, *NPJ Comput Mater* **3**, 46 (2017).
- [6] ELIXIR Research Data Management Kit (RDMkit), *Data Life Cycle: Sharing*, <https://rdmkit.elixir-europe.org/sharing#what-should-be-considered-for-data-sharing>.
- [7] K. Briney, *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success* (Exeter: Pelagic Publishing, UK, 2015).
- [8] The Deutsche Forschungsgemeinschaft (DFG), *Specification of Requirements Relating to the Handling of Research Data in Funding Proposals*, https://www.dfg.de/en/research_funding/announcements_proposals/2022/info_wissenschaft_22_25/.
- [9] Science Europe, *Practical Guide to the International Alignment of Research Data Management*, 2018.
- [10] The Deutsche Forschungsgemeinschaft (DFG), *Handling of Research Data: Checklist for Planning and Description of Handling of Research Data in Research Projects*, 2021.
- [11] re3data.org, *Registry of Research Data Repositories*, <https://doi.org/10.17616/R3D>.

Embedding the de.NBI Cloud in the National Research Data Infrastructure Activities

Nils Hoffmann¹[\[https://orcid.org/0000-0002-6540-6875\]](https://orcid.org/0000-0002-6540-6875), Irena Maus¹[\[https://orcid.org/0000-0003-3335-9514\]](https://orcid.org/0000-0003-3335-9514), Sebastian Beier²[\[https://orcid.org/0000-0002-2177-8781\]](https://orcid.org/0000-0002-2177-8781), Peter Belmann¹[\[https://orcid.org/0000-0002-1294-2869\]](https://orcid.org/0000-0002-1294-2869), Jan Krüger¹, Andreas Tauch¹, The de.NBI Cloud Consortium³, Alexander Sczyrba⁴[\[https://orcid.org/0000-0002-4405-3847\]](https://orcid.org/0000-0002-4405-3847)

¹ Institute of Bio- and Geosciences, Computational Metagenomics (IBG-5), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

² Institute of Bio- and Geosciences, Bioinformatics (IBG-4), Forschungszentrum Jülich GmbH, Jülich, 52425, Germany

³ The de.NBI Cloud Consortium in Germany: Alexander Goesmann¹[\[https://orcid.org/0000-0002-7086-2568\]](https://orcid.org/0000-0002-7086-2568), Justus-Liebig-University Giessen; Roland Eils¹[\[https://orcid.org/0000-0002-0034-4036\]](https://orcid.org/0000-0002-0034-4036), BIH-Zentrum Digitale Gesundheit, Charité - Universitätsmedizin Berlin; Peer Bork¹[\[https://orcid.org/0000-0002-2627-833X\]](https://orcid.org/0000-0002-2627-833X), European Molecular Biology Laboratory Heidelberg; Oliver Kohlbacher¹[\[https://orcid.org/0000-0003-1739-4598\]](https://orcid.org/0000-0003-1739-4598), Eberhard Karls University Tübingen; Ursula Kummer¹[\[https://orcid.org/0000-0002-6413-2382\]](https://orcid.org/0000-0002-6413-2382), BioQuant & Heidelberg University; Rolf Backofen¹[\[https://orcid.org/0000-0001-8231-3323\]](https://orcid.org/0000-0001-8231-3323), Albert-Ludwigs University Freiburg; Ivo Buchhalter¹[\[https://orcid.org/0000-0003-0764-5832\]](https://orcid.org/0000-0003-0764-5832), Deutsches Krebsforschungszentrum Heidelberg; Alexander Sczyrba, Bielefeld University

⁴ Faculty of Technology, Bielefeld University, Universitätsstr. 25, 33615 Bielefeld, Germany

Keywords: Cloud Computing, Federated Research Infrastructure, AAI, Virtual Research Environments, Training

Abstract

In recent years, modern life sciences research underwent a rapid development driven mainly by the technical improvements in analytical areas leading to miniaturization, parallelization, and high throughput processing of biological samples. This has led to the generation of huge amounts of experimental data. To meet these rising demands, the German Network for Bioinformatics Infrastructure (de.NBI)[1] was established in 2015 as a national bioinformatics consortium aiming to provide high quality bioinformatics services, comprehensive training[2], powerful computing capacities (de.NBI Cloud) as well as connections to the European Life Science Infrastructure ELIXIR[3], with the goal to assist researchers in exploring and exploiting data more effectively.

The establishment of the de.NBI Cloud has proven to be a flagship for the de.NBI network (see Fig. 1 for details). It consists of eight federated cloud locations that implement a common governance and use the project application and management workflow provided by the de.NBI Cloud portal. This governance facilitates secure operations of each cloud location through the centralized organization of ISO27001 Information Security Management System (ISMS) training and certification courses for the cloud staff, leading to the progressive certification of our cloud sites. Registration, project resource application and authentication are facilitated by the integration of the LifeScience AAI[4], [5] as an EduGAIN-compatible single sign-on provider, backed by institutional ID providers of universities and research institutes.

Since its foundation, de.NBI Cloud has formed the scientific and collaborative backbone for new major German initiatives like NFDI[6] or EOSC-Life in the European sector of compu-

tational biosciences. Above all, the cooperation with various NFDI consortia such as NFDI4Biodiversity, DataPLANT, GHGA, FAIRagro or NFDI4Microbiota showcases the power, range and flexibility of the de.NBI Cloud, especially for the national life science community.

The de.NBI Cloud portfolio includes several project types designed to suit different use cases and users with varying levels of knowledge in cloud computing. Two project types, OpenStack and Kubernetes, offer maximum flexibility in terms of the configuration of cloud-specific components and allow the installation of any large-scale analysis, stream processing or orchestration framework available in the cloud ecosystem. Both project types are ideal for science gateway developers to offer bioinformatics services to the national and international life sciences communities.

To be more precise, OpenStack makes controlling large pools of computing, storage, and network resources simple. Any interaction with OpenStack can be performed through its dashboard or automated through its API, using well-known cloud infrastructure management frameworks like Terraform or Ansible. While OpenStack enables the provisioning of virtual machines, networks and storage, Kubernetes (K8s) is the state-of-the-art technology for the deployment, scaling, and orchestration of highly available containerized applications. The de.NBI Cloud supports "vanilla" Kubernetes clusters on top of OpenStack using Kubermatic. Features like self-healing, automated rollback, and horizontal scaling make K8s the ideal basis for services. Confidential processing of sensitive data, e.g., pseudonymized patient-related data, is also possible at specific de.NBI Cloud locations, where data security is enforced through separated, secure processing environments. From a legal perspective, this is mediated through data processing agreements between a project's principal investigator and the hosting cloud site.

Our in-house developed project type SimpleVM enables our users to use cloud resources with little to no background knowledge in cloud computing. SimpleVM is an abstraction layer on top of OpenStack to manage single virtual machines (VMs) or clusters thereof. It was designed to support the combination of resources from independent OpenStack installations, thus operating as a multi-cloud platform which is accessible from a single web-based control panel. The entire software stack only requires access to the OpenStack API and can be deployed on any vanilla OpenStack installation. In general, SimpleVM primarily eases the creation and management of individual pre-configured virtual machines and provides web-based access to popular research and development environments such as Rstudio, Guacamole Remote Desktop, Theia IDE, JupyterLab and Visual Studio Code. On top of this functionality, a dedicated mode that simplifies the setup of cloud-based workshops for teaching purposes is also provided. Further, with SimpleVM, de.NBI Cloud users can effortlessly configure and manage their own SLURM-based BiBiGrid clusters with just a few clicks. This feature addresses the needs of researchers who want to run their tools or entire workflows across multiple machines.

For users who want to be able to define data processing workflows from tools available in BioConda and the Galaxy ToolShed with a graphical user interface, the de.NBI cloud infrastructure also hosts the Galaxy[7] service available at usegalaxy.eu. Galaxy also simplifies the discovery and adaptation of existing workflows, that were shared by other users, from multiple scientific domains and enables their execution at scale in the cloud.

In conclusion, the de.NBI Cloud provides the ability to unlock the full potential of research data and enables easier collaboration across different ecosystems and research areas, which in turn enables scientists to innovate and scale-up their data-driven research, not only in the life and computational biosciences, but across the different science domains addressed by the NFDI.

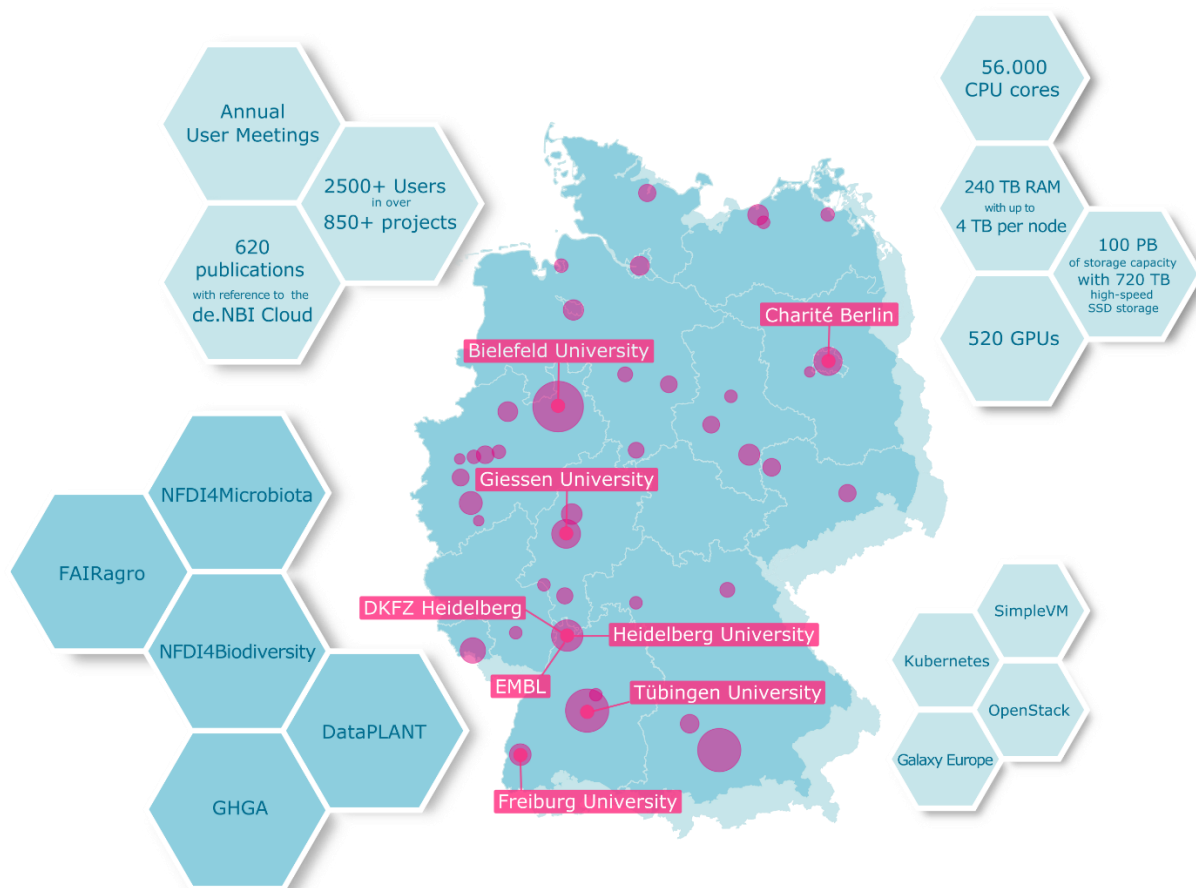


Figure 1. The de.NBI Cloud federation, maintained by the eight German cloud centers in Berlin, Bielefeld, Freiburg, Gießen, Heidelberg and Tübingen. Red circles show the distribution and approximate number of de.NBI Cloud AAI users. The upper left hexagons show data on use and visibility of the cloud, while the upper right hexagons show the total number of provided computing resources. NFDI projects that use cloud resources are listed at the bottom left. The de.NBI Cloud project types, tailored to different use cases and requirements are shown at the bottom right.

Data availability statement

This submission is not based on data.

Author contributions

NH, IM, SB, PB, JK, AT and AS drafted and wrote the original draft, all authors reviewed, edited and approved the draft manuscript. AG, RE, PB, OK, UK, RB, IB and AS acquired the funding for the de.NBI cloud and are, together with AT, responsible for project administration and supervision.

Competing interests

The authors declare that they have no competing interests.

Funding

We acknowledge funding by the German Ministry for Education and Research (BMBF) under the grant numbers (FKZ) 031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, and 031A537D.

Acknowledgement

Operating a federated infrastructure like the de.NBI Cloud is not possible with a small group of people alone. The authors would therefore like to thank all cloud site personnel and people involved with development and operations, past and present, for their dedication and support to provide an excellent service to the life sciences community in Germany.

References

- [1] A. Pühler, "Bioinformatics solutions for big data analysis in life sciences presented by the German network for bioinformatics infrastructure," *J Biotechnol*, vol. 261, p. 1, Nov. 2017, doi: 10.1016/j.jbiotec.2017.08.025.
- [2] D. Wibberg *et al.*, "The de.NBI / ELIXIR-DE training platform - Bioinformatics training in Germany and across Europe within ELIXIR," *F1000 Research*, vol. 8, p. 1877, Nov. 2019, doi: 10.12688/f1000research.20244.1.
- [3] A. Tauch and A. Al-Dilaimi, "Bioinformatics in Germany: toward a national-level infrastructure," *Brief Bioinform*, vol. 20, no. 2, pp. 370–374, Mar. 2019, doi: 10.1093/bib/bbx040.
- [4] M. Linden *et al.*, "Common ELIXIR Service for Researcher Authentication and Authorisation," *F1000Res*, vol. 7, p. ELIXIR-1199, 2018, doi: 10.12688/f1000research.15161.1.
- [5] P. Belmann *et al.*, "de.NBI Cloud federation through ELIXIR AAI," *F1000Res*, vol. 8, p. 842, 2019, doi: 10.12688/f1000research.19013.1.
- [6] S. Kraft *et al.*, "Nationale Forschungsdateninfrastruktur (NFDI) e. V.: Aufbau und Ziele," *BFDM*, no. 2, pp. 1–9, Jul. 2021, doi: 10.17192/bfdm.2021.2.8332.
- [7] The Galaxy Community, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update," *Nucleic Acids Research*, vol. 50, no. W1, pp. W345–W351, Jul. 2022, doi: 10.1093/nar/gkac247.

Monitoring the state of open and FAIR data in Helmholtz

A data-harvesting and dashboard-approach by HMC

Gabriel Preuß ¹²[\[https://orcid.org/0000-0002-3968-2446\]](https://orcid.org/0000-0002-3968-2446),
Alexander M. Schmidt ¹²[\[https://orcid.org/0009-0005-1368-6114\]](https://orcid.org/0009-0005-1368-6114),
Mojeeb R. Sedeqi ¹²[\[https://orcid.org/0000-0002-9694-0122\]](https://orcid.org/0000-0002-9694-0122),
Vivien Serve ¹²[\[https://orcid.org/0000-0001-9603-7630\]](https://orcid.org/0000-0001-9603-7630),
Oonagh Mannix ¹²[\[https://orcid.org/0000-0003-0575-2853\]](https://orcid.org/0000-0003-0575-2853), and
Markus Kubin ¹²[\[https://orcid.org/0000-0002-2209-9385\]](https://orcid.org/0000-0002-2209-9385)

¹Helmholtz Metadata Collaboration (HMC), Hub Matter

²Helmholtz-Zentrum Berlin für Materialien und Energie, Germany

Abstract: In this contribution we present an integrated approach to monitoring and assessing the state of open and FAIR data in the Helmholtz Association. The project is part of a multi-method approach by Hub Matter in the Helmholtz Metadata Collaboration (HMC).

In a harvesting-approach, data published by Helmholtz researchers is found starting from literature metadata, harvested from the research centers. Data publications linked to that literature are identified using the SCHOLIX API. In a first approach to automated FAIR assessment, we adopted the F-UJI framework, as developed by the FAIRsFAIR consortium.

The information collected is presented in an interactive dashboard. It allows to explore in which repositories Helmholtz researchers make their data publicly available, to engage Helmholtz communities, and to identify gaps towards improving the FAIRness of Helmholtz data.

The dashboard is publicly available on <https://fairdashboard.helmholtz-metadaten.de>. The general approach as well as all program code are reusable by all research communities.

Keywords: OPEN DATA, FAIR DATA, METADATA HARVESTING, DASHBOARD

1 Introduction

The Helmholtz Metadata Collaboration (HMC) platform was launched in late 2019 to turn FAIR (Findable, Accessible, Interoperable, Reusable)[1] data practices within the Helmholtz Association into reality. By leveraging the visibility and re-usability of data the HMC platform aims to develop and consolidate community-expertise in metadata across Helmholtz.

To develop effective strategies, key information about the state of FAIR data practices is required:

1. Where (in which repositories) and how much is Helmholtz data published?
2. How can we identify and analyze gaps in the FAIRness of this data?

Answering these questions will help us monitor the data publishing landscape in Helmholtz and to identify action items towards a FAIR data space in Helmholtz.

2 How can we find Helmholtz data? An approach to data-harvesting and automated FAIR assessment

Being interested in finding data publications by Helmholtz researchers, we started by collecting literature metadata from Helmholtz libraries. This is done via the *OAI Protocol for Metadata Harvesting* (OAI-PMH)[2]. Data publications are often published together with them, so we can benefit from their well curated and rich metadata. The connection between literature publications and their related data publications is done via the external tool *ScholarXplorer*[3], which finds related data publications for literature publications by a DOI and offers structured metadata about them. *ScholarXplorer* allows to quickly identify and gather information about related data publications that may not be immediately apparent through other means. Finally, the toolbox evaluates data publications using F-UJI and provides a FAIR score. F-UJI [4], [5] is a web service that programmatically assesses the FAIRness of research data objects based on metrics developed by the FAIRsFAIR project[6].

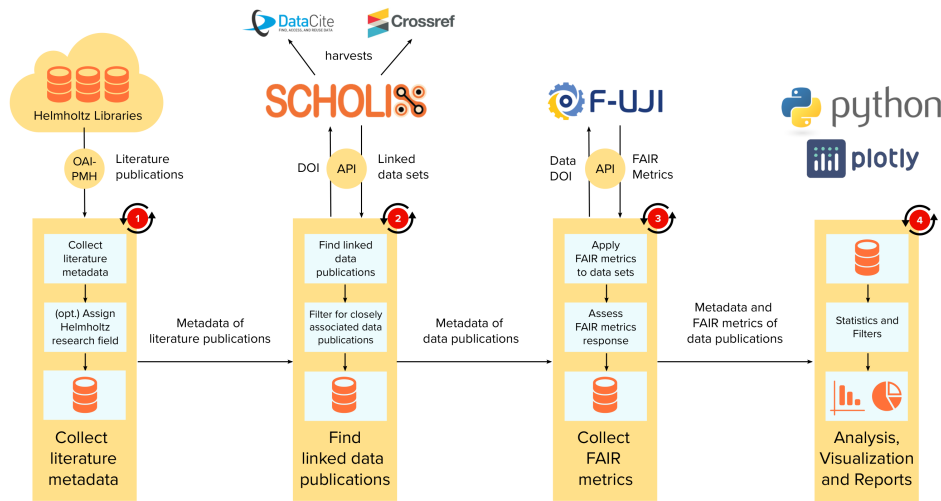


Figure 1. Overview Toolbox Workflow

2.1 HMC Toolbox for Data-Harvesting

The implementation of the *HMC Toolbox* was done in *Python3*[7] and is based on five independently maintainable modules; named *Harvester*, *Metadata Extractor*, *Linked Data Finder*, *FAIR Meter* and the *Exporter*.

1. **Harvester**: the Harvester module requests a given Helmholtz library API to get information about available literature publications. We implemented harvesters for OAI-PMH APIs via the well known libraries *OAI-PMH Harvest*[8] and *Sickle: OAI-PMH for Humans*[9]. Also, a harvester for plain file downloading is offered due to the fact that some

- libraries offer their data in CSV files. The Harvester module offers the option to be extended by other harvester types easily.
2. Metadata Extractor: the libraries offer their data in many formats and schemas hence some kind of mapping needs to be done. Because the Dublin Core Metadata Schema is a must-have for OAI-PMH APIs the toolbox brings a mapping for that by default as well as for the old but still widely used marcxml schema.
 3. Linked Data Finder: within this module the extracted metadata is enriched with additional information which can come from any implemented source. By default, the mentioned *ScholarXplorer*[3] is used to enrich a literature publication which its data publications. But also any other source is conceivable here.
 4. FAIR Meter: if datasets are found for a literature publication they are successively evaluated for their compliance with the FAIR principles using F-UJI[4] as a first approach. We employ the F-UJI docker container offered on GitHub[10] to set up a local F-UJI server.
 5. Exporter: to ensure re-usability of the collected data we export all metadata harvested in the *JSON* format. The Exporter module offers standard output options like screen or file output and is easily extendable to other output targets like databases.

3 How can we engage communities? An interactive dashboard approach

The information collected is presented in an interactive dashboard. It allows to explore in which repositories Helmholtz researchers make their data publicly available, to engage Helmholtz communities, and to identify gaps towards improving the FAIRness of Helmholtz data.

3.1 HMC FAIR Data Dashboard Implementation

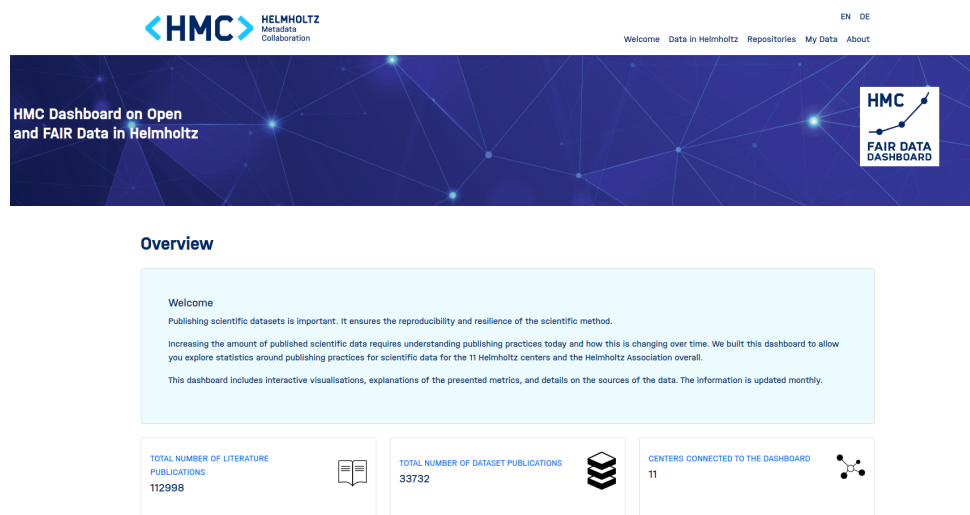


Figure 2. Welcome Page HMC FAIR Data Dashboard

We decided to use Dash[11] together with Flask[12] as a framework to render our plots on the dashboard.

The data collected with the *HMC Toolbox* was stored in a database and delivered to the dashboard. All plots are interactive, allowing users to filter and focus on desired

issues. Additionally an option is offered to check the FAIR score of publications in the database.

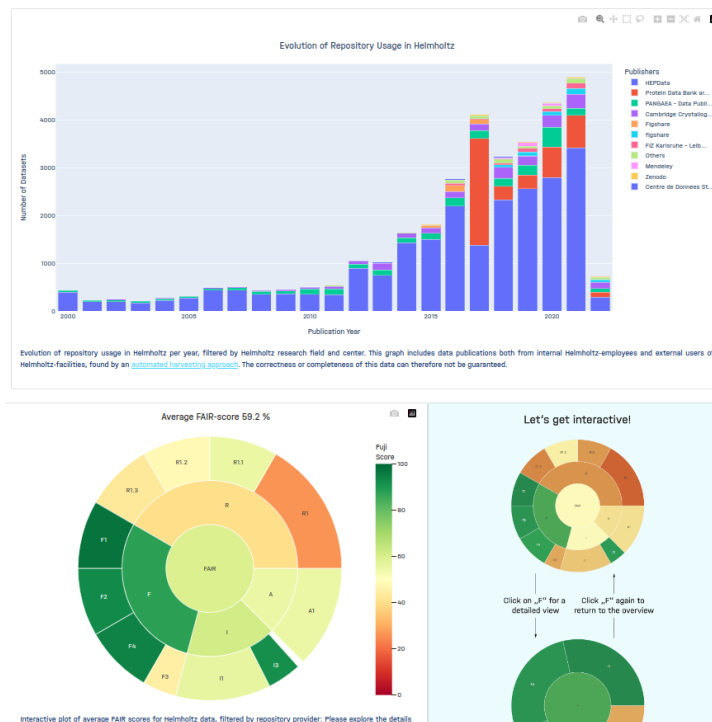


Figure 3. Screenshot of interactive dashboard figures

Automated approaches are limited and depend heavily on the data quality of their sources. On the other hand, the goal is to improve data quality from the FAIR perspective in the research environment. This is a mutually reinforcing loop — as the state of data is publicly communicated by the dashboard, the data producers feel empowered to improve their data which improves the quality of the harvested data for the dashboard.

The used FAIR scoring method used by F-UJI is just one perspective on the FAIRness of data publications. Other methods need to be added to the FAIR meter and evaluate the data in different ways. Additionally counting data publications is also sensitive as it can be challenging to determine what counts as a data publication. Finally there is a lack of standardized (meta-)data quality, which can make it difficult to compare results across different sources.

4 Conclusions & Outlook

We are committed to improving the data-harvesting and dashboard approaches. As a first step, the code is publicly available (see data availability statement) so that other researchers and institutions can use and build upon it. More Helmholtz centers will be integrated in later updates.

In addition, the modular approach allows for easy incorporation of other scoring metrics and harvesting options.

In conclusion, the presented approach represents an important step towards achieving a more FAIR data publishing environment in the Helmholtz research environments. By monitoring data publishing practices, and providing analyses of the data, HMC is working towards a FAIR data space in the Helmholtz Association.

Data availability statement

The HMC Dashboard on Open and FAIR data in Helmholtz is published on <https://fairdashboard.helmholtz-metadaten.de/>.

All program code, both for the harvesting toolbox and the interactive dashboard are made publicly available and reusable in the following [GitLab repository](#).

Competing interests

The authors declare that they have no competing interests.

Funding

This publication was supported within the hub Matter at the Helmholtz-Zentrum Berlin by the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative.

Acknowledgements

We thank the Helmholtz-Zentrum Berlin for the necessary computing resources to implement that project.

References

- [1] M. Wilkinson, M. Dumontier, I. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Sci Data* 3, 160018, 2016. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
- [2] C. Lagoze, H. V. de Sompel, M. Nelson, and S. Warner. "The open archives initiative protocol for metadata harvesting - v.2.0." (2015), [Online]. Available: <https://www.openarchives.org/OAI/openarchivesprotocol.html> (visited on 04/25/2023).
- [3] A. Burton, H. Koers, P. Manghi, M. Stocker, *et al.*, "The scholix framework for interoperability in data-literature information exchange," *D-Lib Magazine*, vol. 23, 2017. DOI: [10.1045/january2017-burton](https://doi.org/10.1045/january2017-burton). [Online]. Available: <https://www.dlib.org/dlib/january17/burton/01burton.html>.

- [4] A. Devaraju and R. Huber, *F-uji - an automated fair data assessment tool*, version v1.0.0, Oct. 2020. DOI: [10.5281/zenodo.4063720](https://doi.org/10.5281/zenodo.4063720). [Online]. Available: <https://doi.org/10.5281/zenodo.4063720>.
- [5] A. Devaraju and R. Huber, "An automated solution for measuring the progress toward fair research data," *Patterns*, vol. 2(11), 2021. DOI: [10.1016/j.patter.2021.100370](https://doi.org/10.1016/j.patter.2021.100370).
- [6] A. Devaraju, R. Huber, M. Mokrane, et al., *Fairsfair data object assessment metrics*, version 0.5, Apr. 2022. DOI: [10.5281/zenodo.6461229](https://doi.org/10.5281/zenodo.6461229). [Online]. Available: <https://doi.org/10.5281/zenodo.6461229>.
- [7] "Python.org, The official home of the python programming language." (), [Online]. Available: <https://www.python.org/> (visited on 04/25/2023).
- [8] "Github, Bloomonkey/oai-harvest: Python package for harvesting records from oai-pmh provider(s)." (), [Online]. Available: <https://github.com/bloomonkey/oai-harvest> (visited on 04/25/2023).
- [9] "Github, Mloesch/sickle: Sickle: Oai-pmh for humans." (), [Online]. Available: <https://github.com/mloesch/sickle> (visited on 04/25/2023).
- [10] "Github, Pangaea-data-publisher/fuji: Fairsfair research data object assessment service." (), [Online]. Available: <https://github.com/pangaea-data-publisher/fuji> (visited on 04/25/2023).
- [11] "Github, Plotly/dash: Data apps & dashboards for python. no javascript required." (), [Online]. Available: <https://github.com/plotly/dash> (visited on 04/25/2023).
- [12] "Github, Pallets/flask: The python micro framework for building web applications." (), [Online]. Available: <https://github.com/pallets/flask> (visited on 04/25/2023).

NFDI4DS Gateway and Portal

Ricardo Usbeck¹[\[https://orcid.org/0000-0002-0191-7211\]](https://orcid.org/0000-0002-0191-7211), Tilahun Abedissa Taffa¹[\[https://orcid.org/0000-0002-2476-8335\]](https://orcid.org/0000-0002-2476-8335), Rudy Alexandro Garrido Veliz¹[\[https://orcid.org/0009-0008-1582-2417\]](https://orcid.org/0009-0008-1582-2417), Rana Abdullah¹[\[https://orcid.org/0009-0000-2652-5129\]](https://orcid.org/0009-0000-2652-5129), Najeebullah Shams¹[\[https://orcid.org/0000-0003-3725-2554\]](https://orcid.org/0000-0003-3725-2554), Bianca Wentzel²[\[https://orcid.org/0000-0002-9218-5676\]](https://orcid.org/0000-0002-9218-5676), Zongxiong Chen²[\[https://orcid.org/0000-0003-2452-0572\]](https://orcid.org/0000-0003-2452-0572), and Sonja Schimmler²[\[https://orcid.org/0000-0002-8786-7250\]](https://orcid.org/0000-0002-8786-7250)

¹University of Hamburg, Germany

²Fraunhofer FOKUS, Germany

Abstract: NFDI4DataScience (NFDI4DS) is a consortium to support researchers in all stages of the research data lifecycle to conduct their research in line with the FAIR principles. The developed infrastructure targets researchers from a wide range of disciplines in data science and AI.

We present the ideas of the NFDI4DS gateway and the NFDI4DS portal. Two approaches to navigate digital objects (articles, data, machine learning models, workflows, scripts/code, etc.) from various NFDI4DS resources such as the ORKG, the DBLP database, and other research knowledge graphs (KGs). Transparency, reproducibility, and fairness will be fostered by a step-wise integration of existing and newly developed services into the overall system.

With this paper, we want to engage with the community and understand the needs and challenges of researchers in various disciplines regarding data science and AI. Therefore, we will discuss the currently developed prototypes and outline our plans for future development steps.

Keywords: Enabling RDM (incl. software), Linking RDM, FAIR, FDO

1 Introduction

NFDI4DS Gateway and Portal

The current paradigm shift towards data-driven and deep learning-based approaches in data science requires an expert-level understanding of available resources. Consequently, the community needs an integrated gateway and portal functioning as a search engine over multiple scholarly repositories and services. Through a unified search and exploration interface, users will be able to query a wide range of scientific databases. The results are currently mapped to Schema.org¹ and DCAT-AP². The NFDI4DS gateway offers querying in an ad-hoc fashion, the NFDI4DS portal provides a harvesting-based approach to account for larger research data dumps.

¹<https://schema.org>

²<https://github.com/SEMIGeu/DCAT-AP/>

2 Approach

Architecture

The NFDI4DS gateway receives a search key, executes it against the repositories using their respective APIs, and outputs human-readable results. The heterogeneous API results are mapped to their respective schema. Since each entity has a different identifier in each repository, the background controller deduplicates results. In contrast, the NFDI4DS portal filters the underlying knowledge base of harvested metadata based on filters or by keyword utilizing the integrated search engine.

Currently, the ad-hoc based approach³ is utilizing ten open-source scholarly repositories, namely DBLP⁴, OpenAlex⁵, CORDIS⁶, European Language Grid⁷, GEPRIS⁸, GESIS⁹, ORCID¹⁰, RESODATE¹¹, WIKIDATA¹², IEEE¹³, and Zenodo¹⁴. Among these repositories, DBLP, OpenAlex, IEEE, GESIS, RESODATE, WIKIDATA, and Zenodo provide research resources like publications, datasets, software, etc. GEPRIS provides Deutsche Forschungsgemeinschaft (DFG) funded projects, likewise, CORDIS is a primary source for projects financed by the European Union (EU) commission. ELG is a platform to avail multi-lingual, cross-lingual, and mono-lingual language technologies in the EU. Unlike the others, to distinguish researchers uniquely, ORCID provides a unique persistent researcher-owned and controlled digital identifier.

Similar to the gateway, the harvesting-based approach¹⁵ retrieves metadata via APIs and harvesting interfaces of different repositories. The system is based on Piveau[1], a fully-fledged data management ecosystem. During the harvesting process, the metadata is transformed into the machine-readable format RDF using the DCAT-AP specification, which is based on W3C's Data Catalogue Vocabulary (DCAT) providing a plethora of properties, vocabularies and guidelines to express information about Open Data. This data is stored, indexed, and made available via a separate frontend and via APIs. In addition, the portal provides a SPARQL endpoint that enables direct querying of the underlying knowledge graph formed by the harvested metadata.

The plan is to integrate both systems so that the data from both approaches can be accessed via one entry point. In the future, we will not only be an aggregator but also run additional services like an assessment service. We also foresee using the RDF graphs for further downstream tasks as well as offering the architecture to other consortia.

Search Paradigms

Searching paradigms can be classified as keyword, structured/controlled, or natural language questions, depending on the technique used to process a given query. Key-

³<https://nfdi-search.nliwod.org/>

⁴<https://dblp.org>

⁵<https://docs.openalex.org>

⁶<https://cordis.europa.eu>

⁷<https://live.european-language-grid.eu>

⁸<https://gepris.dfg.de/gepris/OCTOPUS>

⁹<https://www.gesis.org/home>

¹⁰<https://orcid.org>

¹¹<https://resodate.org/resources/>

¹²https://www.wikidata.org/wiki/Wikidata:Main_Page

¹³<https://www.ieee.org>

¹⁴<https://zenodo.org>

¹⁵<https://meta4ds.fokus.fraunhofer.de>

word searches match only lexical terms and do not consider structural or semantic mappings. To enhance precision, the controlled keyword search paradigm was developed. Structured searches require users to write a structured query like SQL or SPARQL and should be familiar with the querying interfaces. In contrast, the Question and Answering (QA) searching paradigm enables users to input natural language questions and receive answers by analyzing and reasoning over the underlying data source. The NFDI4DS gateway currently only utilizes keyword matching but foresees providing a Large Language Model-powered chatbot for search result page analysis. The NFDI4DS portal is based on keyword and faceted search. The faceted search enables users to refine search results by applying filters based on different attributes, enhancing the search experience and facilitating the exploration of specific subsets of data.

Frontend

Both systems provide a frontend showcasing the data in a direct way (see Figure 1). The interfaces are composed of a header menu as well as a search bar for exploring the data. The gateway menu provides links to events, the community and additional services¹⁶ while the portal menu also includes a link to the SPARQL endpoint of the system. Both systems show a list of obtained results with basic information divided into tabs based on their content. While the portal currently only supports datasets and catalogues, the gateway provides a multitude of tabs covering researchers, articles or events. Each listed result links to a details page containing more detailed information on the data as well as links to additional NFDI4DS services. The NFDI4DS portal also provides links to the underlying linked semantic data. The gateway will enable searching via a chatbot through the user interface utilizing a different approach than the portal which offers filtering of results based on facets.

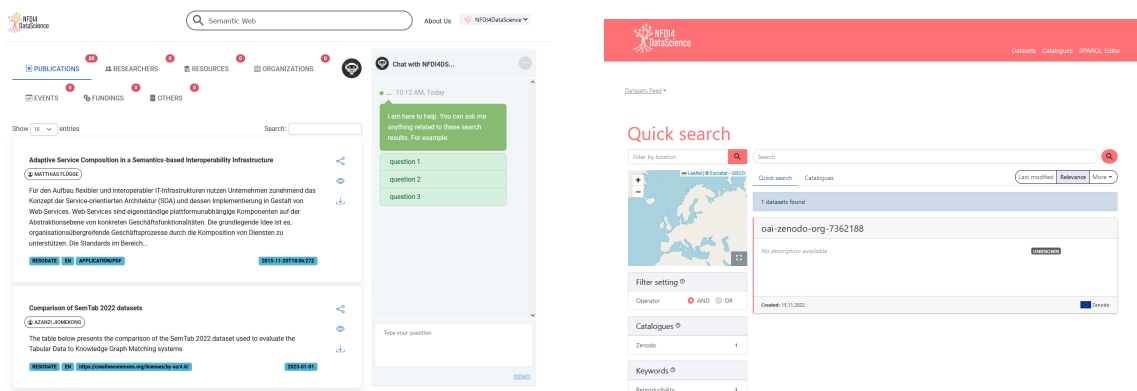


Figure 1. Interface of the gateway prototype (left) and the portal prototype (right).

3 Service Integration

The NFDI4DS gateway and portal integrate or will integrate different services in the future. Key to the success is a seamless integration of these services. In the following, we describe some central integrations that are currently under development:

¹⁶Using <https://gitlab.com/TIBHannover/nfdi4ds/nfdi4ds-widget>

NFDI4DS Research Knowledge Graph

NFDI4DS aims at providing well-structured public knowledge about scholarly resources and their adoption and relations to facilitate various use cases. The NFDI4DS research KG will entail automatically extracted metadata about resource relations, e.g. software mentions in scholarly publications, highly quality-controlled manual annotations of scholarly resources, and community annotations of scholarly publications from the ORKG.

NFDI4DS Registries

The consortium aims at providing registries for different digital objects, one of which is the DBLP computer science bibliography. DBLP and the ORKG started linking author and publication entities within their respective knowledge graphs. Using this linkage, ORKG can make use of DBLP's semantic organization and intellectual author disambiguation which, among other information, provides author identities even in the absence of schemes like ORCID [2]. At the same time, DBLP will make use of ORKG's deep semantic description of research content to enhance the information given on DBLP's website.

4 Conclusion

The NFDI4DS gateway and portal is currently under heavy development. Some features will be developed only after the finalization of this paper. Since NFDI is on the move, it is considered, that with this submission, the need to get in touch with and understand data scientists across consortia and beyond, can be satisfied.

Competing interests

The authors declare that they have no competing interests.

Funding

This joint project received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number: NFDI4DataScience (460234259).

References

- [1] F. Kirstein, K. Stefanidis, B. Dittwald, S. Dutkowski, S. Urbanek, and M. Hauswirth, "Piveau: A large-scale open data management platform based on semantic web technologies," in *The Semantic Web*, A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, *et al.*, Eds., Cham: Springer International Publishing, 2020, pp. 648–664, ISBN: 978-3-030-49461-2.
- [2] L. L. Haak, M. Fenner, L. Paglione, E. Pentz, and H. Ratner, "ORCID: a system to uniquely identify researchers," *Learn. Publ.*, vol. 25, no. 4, pp. 259–264, 2012. DOI: [10.1087/20120404](https://doi.org/10.1087/20120404). [Online]. Available: <https://doi.org/10.1087/20120404>.

Establishing the Research Data Management Container in NFDIxCS

Firas Al Laban¹[\[https://orcid.org/0000-0001-8072-9384\]](https://orcid.org/0000-0001-8072-9384), Jan Bernoth¹[\[https://orcid.org/0000-0002-4127-0053\]](https://orcid.org/0000-0002-4127-0053),
Michael Goedicke², Ulrike Lucke¹[\[https://orcid.org/0000-0003-4049-8088\]](https://orcid.org/0000-0003-4049-8088),
Michael Striewe²[\[https://orcid.org/0000-0001-8866-6971\]](https://orcid.org/0000-0001-8866-6971), Philipp Wieder³[\[https://orcid.org/0000-0002-6992-1866\]](https://orcid.org/0000-0002-6992-1866), and
Ramin Yahyapour³[\[https://orcid.org/0000-0002-9057-4395\]](https://orcid.org/0000-0002-9057-4395)

¹ Universität Potsdam, Germany

² University of Duisburg-Essen, Germany

³ Gesellschaft für Wissenschaftliche Datenverarbeitung mbH, Göttingen, Germany

Keywords: Research Data Management Container, NFDIxCS, Roadmap, Research Software, Software Sustainability, FAIR

NFDIxCS¹ is a consortium within the family of NFDI², which defines and establishes a research data management (RDM) infrastructure for Computer Science (CS). Based on a broad community process the various types of research data, their metadata and quality criteria are agreed upon in the community. The resulting research data, along with all associated supplementary information and context such as software, metadata, and the corresponding execution environment, are provided as an integral part of the overall infrastructure to meet the FAIR principles [1].

One key aspect of this infrastructure, which encapsulates connected research artifacts into a package object format, is the Research Data Management Container (RDMC) [2]. A central question addressed by the RDMC is how software artifacts can be sustainably preserved in the scientific system for a long time. The citation of software and data reveals a tension between the interests of scientists and publishers [3]; however, after archiving the software and providing a traceable citation, it is still not guaranteed that the addressed findings can be executed. The goal pursued with the RDMC is to create a container that links data, software, and execution environments so that the artifact can be used in the future and without much effort for Research Software Engineers under previously determined access regulations.

¹ <https://www.nfdixcs.org>

² <https://www.nfdi.de>, <https://www.dfg.de/nfdi>

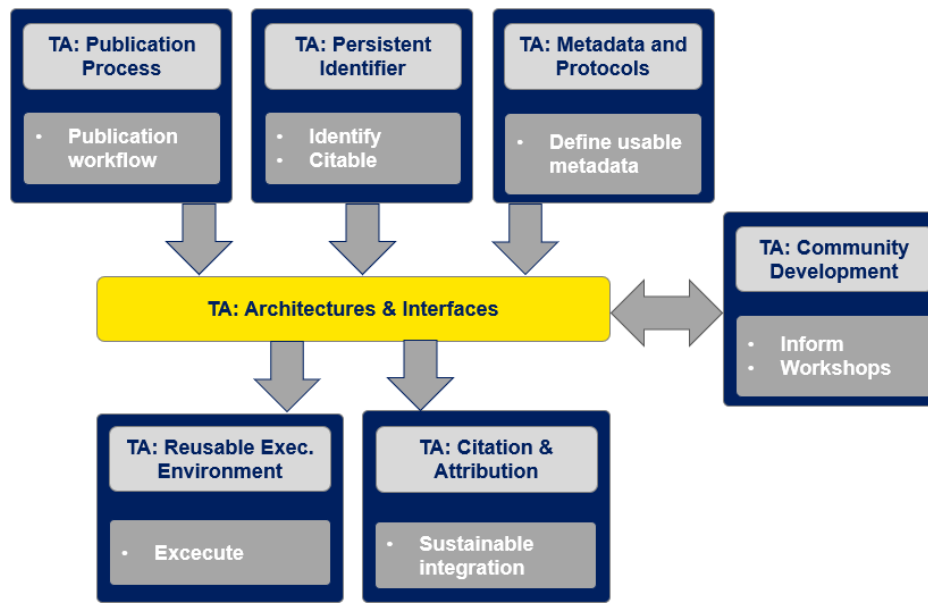


Figure 1. Role of the TA *Architectures & Interfaces* within the NFDIxCS Consortium.

The conception and development of the RDMC and its execution platform take place in the Task Area (TA) *Architectures & Interfaces*, which is closely linked to the other TAs (Figure 1).

Input from other TAs such as *Persistent Identifiers* and *Metadata and Protocols* is crucial for making containers findable and accessible in line with the FAIR principles. The TA *Publication Process* enables the RDMC to integrate automatic mechanisms in its workflow. Additionally, gathering requirements, conducting workshops for implementation, and community feedback are collected in the TA *Community Development* to integrate the user community into the RDMC development process. The deployment of the RDMC in an execution environment is addressed in the TA *Reusable Execution Environment*. The TA *Citation and Attribution* plays a vital role in integrating the RDMC into publication processes, ensuring proper credit and recognition for all involved parties.

The following roadmap (Figure 2) outlines stages for the development of the RDMC and its platform. Additionally, it demonstrates how external influences from the NFDIxCS consortium and beyond are considered.

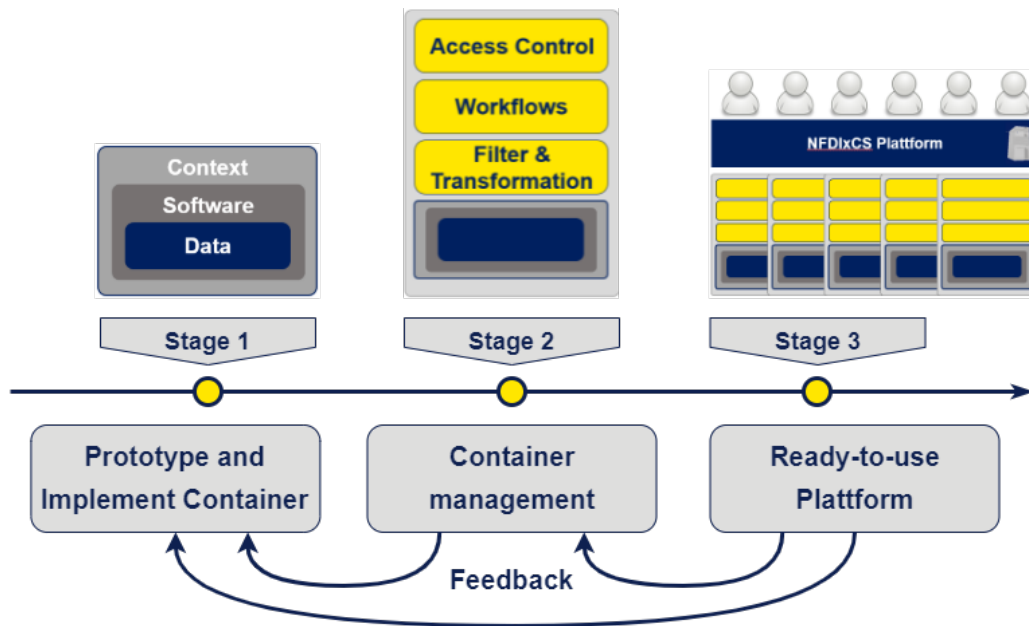


Figure 2. Implementing the RDMC.

Stage 1: Prototype and Implement Container

This stage is concerned with a prototype that realizes the basic requirements. The design and architecture of the RDMC prototype will be mostly agnostic regarding metadata formats, identifier standards, security mechanisms and similar. It will integrate the components that are needed to collect and organize data, and will provide the corresponding interfaces regardless of the data size or data type. One of the main characteristics of RDMC is to be flexible enough to deal with the content with no restricted rules and to use different metadata standards in parallel that may evolve within the NFDI community and beyond.

This stage will define data storage organization within the RDMC in such a way that metadata and related information about the execution environment is readable even if there is no access to the context. It will also support versioning of contents as well as of the related execution environments. Moreover, the RDMC will provide a mechanism to associate metadata and persistent identifiers both with the whole RDMC instance and specific data within that instance, e.g. for citation and attribution. The mechanism must be open to support different formats both for identifiers and metadata (such as Citation.cff [4] and DOI). This has also an international dimension, considering e.g. EOSC³ for interoperable metadata.

Stage 2: Container management

Access control, workflows, and filter & transformation are specific features integrated into the RDMC. They control the access to and the processing of the data. This stage will focus on implementing modules that make these features usable based on an executable container platform like Docker⁴. Again, the challenge is to keep the definitions open and to not impose unnecessary restrictions, e.g. by making features only available within a specific platform implementation or a self-contained implementation for executable RDMCs.

One of the most important modules is the RDMC access layer that restricts access to the container's contents according to a user's permissions. It needs to allow for a fine-grained access control to make sure that permissions can be associated with any identifiable object

³ <https://eosc-portal.eu/>

⁴ <https://www.docker.com/>

within the RDMC, while metadata is universally accessible. Another module will implement support for workflows according to the use cases for RDMC, where the state of the workflow is encapsulated in the container. The latter is important to allow access rights to apply to specific phases of a workflow, such as access for reviewers during the publication process. Finally, filters and transformations as plugin modules may modify data access in specific use cases.

Stage 3: Ready-to-use Platform

By completing this stage, the RDMC will be established and be ready for the end users. The core activity in this stage is to create and launch a platform that allows to create, host, and access RDMCs. The main challenge is the creation of a (potentially federated) infrastructure, while most technical aspects of container management are already tackled in the previous stages. Technical management and monitoring features will also be considered in this stage to launch an evaluation process that collects evidence and factors regarding the best practices of the platform and RDMC in use. This ensures the sustainable development of the NFDIxCS infrastructure beyond the launching phase.

Conclusion

We have outlined the requirements, characteristics, and key concepts of the RDMC, which is the starting point of the NFDIxCS TA *Architectures & Interfaces* for the next 5 years. The RDMC encapsulates all related research data, contextual information including the software packages used, and the execution environment. This makes it much easier to realize the FAIR principles and the reproducibility of results for research data in Computer Science. We also envisage applications outside CS and will collaborate closely with other NFDI consortia.

Disclaimer

Parts of this text could be generated or rephrased by ChatGPT, DeepL Write, LanguageTool, and Google Docs spell checking, but were carefully checked and revised by the authors.

References

1. C. Engelhardt, How to be FAIR with your data. Göttingen: Göttingen University Press, 2022. DOI: <https://doi.org/10.17875/gup2022-1915>.
2. M. Goedicke and U. Lucke, "Research Data Management in Computer Science - NFDIxCS Approach," 2022. In: Demmler, D., Krupka, D. & Federrath, H. (Hrsg.), INFORMATIK 2022. Gesellschaft für Informatik, Bonn. (S. 1317-1328). DOI: [10.18420/inf2022_112](https://doi.org/10.18420/inf2022_112).
3. M. Harrison, "Open Science, Publications and Code", In: OPEN SCIENCE EUROPEAN CONFERENCE, Proceedings of the Paris Open Science European Conference: OSEC 2022. Paris, Marseille, France: OpenEdition Press, 2022, ISBN: 9791036545627, pp. 183-187. DOI: [10.4000/books.oep.15829](https://doi.org/10.4000/books.oep.15829).
4. S. Druskat, J. H. Spaaks, N. Chue Hong, R. Haines, and J. Baker, "Citation File Format (CFF) - Specifications," 2021. DOI: [10.5281/zenodo.4813122](https://doi.org/10.5281/zenodo.4813122)

RO-Crates meets FAIR Digital Objects

Leyla Jael Castro¹[\[https://orcid.org/0000-0003-3986-0510\]](https://orcid.org/0000-0003-3986-0510), Stian Soliland-Reyes²[\[https://orcid.org/0000-0001-9842-9718\]](https://orcid.org/0000-0001-9842-9718),
and Dietrich Rebholz-Schuhmann^{1,3}[\[https://orcid.org/0000-0002-1018-0370\]](https://orcid.org/0000-0002-1018-0370)

1 ZB MED Information Centre for Life Sciences, Cologne, Germany

2 The University of Manchester, Manchester, United Kingdom

3 University of Cologne, Cologne, Germany

Abstract. RO-Crates makes it easier to package research digital objects together with their metadata so both dependencies and context can be captured. Combined with FAIR good practices such as the use of persistent identifiers, inclusion of license, clear object provenance, and adherence to community standards, RO-crates provides a way to increase FAIRness in science. In this abstract we present the use of RO-Crates, combined with Linked Data best practices, as an implementation for lightweight FAIR Digital Objects, and its possible use in NFDI consortia.

Keywords: RO-Crates, FAIR, FAIR Digital Objects

1. Background

Linked Data (LD) [1] builds upon standards for the Web (e.g., Hypertext Transfer Protocol (HTTP), and Unique Resource Identifier (URIs)) and the Semantic Web (e.g., Resource Data Framework (RDF)), making it easier to interconnect resources to each other via meaningful links. LD can be seen as a precursor and motivator [2] of the Findable, Accessible, Interoperable and Reusable FAIR Principles [3] as it already discuss findability via URIs, accessibility using HTTP, interoperability thanks to the use of semantic data, and reusability by providing a license. Built on top of FAIR, the FAIR Digital Objects (FDOs) [4] aims at clearly separating the metadata from the digital object from its possible materializations with an additional operational layer for machines to directly act over the underlying digital object, with all these components forming a cohesive unit.

RO-Crate [5] follows best practices from LD to provide a lightweight approach to package digital objects together with their metadata, making it easier for researchers to capture both dependencies and context. RO-Crate uses structured metadata in JSON-LD and based on Schema.org [6] to define profiles corresponding to different digital objects. RO-Crate profiles encourage the use of FAIR best practices such as using Persistent Identifiers, (PIDs), providing a license, and linking to related objects. Although the use of LD and RO-Crates does not guarantee compliance to FDO specification requirements, their combination together with a set of constraints on the metadata [7] would turn RO-crates into FDOs. For instance, by requesting the declaration of the type in the metadata, assigning a separate PID to metadata, and using HTTP operations in a consistent manner.

In the rest of this abstract we introduce the use of RO-Crates to implement FDOs (building upon [8], discuss some possible uses and benefits in the National Research Data Infrastructure (NFDI) in Germany, and present some future work.

2. RO-Crates Compliance to FDO Requirements

The FDO Forum has published a set of FDO Requirement Specifications which, in their Version 1.0, comprise nine generic guidelines and twelve FDO requirements (FDORs) [9]. Here we will focus on the FDORs. Assuming an RO-crate as an FDO, we will show how RO-crates can fulfill those FDORs.

Every FDO is assigned a PID (**FDOR1**). PIDs are commonly assigned by third-parties, for instance Digital Object Identifiers (DOIs) are governed by the International DOI Foundation [10] while the Permanent Identifiers for the Web (w3id) [11] are managed by the community via Pull Requests to their GitHub repository. PIDs can be assigned to RO-crates to identify the package as a whole. **FDOR2** states that the PID assigned to an FDO resolves to an structured PID record (being a PID record an FDO, in this case an RO-Crate) following a PID Profile (defined by the community). RO-crates support the use of profiles which effectively define the metadata (types and attributes) accompanying the packaged DO. To fulfill **FDOR2**, such metadata should include the DO type. Different RO-Crate profiles correspond to different DO types for which RO-Crates rely on profiles defined by the Bioschemas [12] community (which include the DO type). Once the FDO Typing System is defined, RO-Crate and Bioschemas will need to align to those. The PID Record (here the RO-Crate) includes mandatory and optional FDO attributes and attributes agreed by the community (**FDOR3**); RO-Crates already follow community agreed types and attributes. If a bit-sequence is available for an FDO, it has to be accessible through the FDO (**FDOR4**). RO-Crates are flexible in this sense, they can contain the actual DO file, or point to a location where it is hosted. An additional effort may be required for RO-Crates to check the accessibility of files externally hosted. **FDOR5** refers to accessibility via standard protocols; currently, RO-Crates can be accessed via HTTP.

FDOR6 deals with the availability of Create, Read, Update, Delete (CRUD) operations for FDOs. RO-Crates do not directly support operations. This is an area where an extension, e.g., a supporting API, would be required. **FDOR7** aims at securing the integrity between FDO Types and operations, which should be maintained in registries. We see this as a requirement external to FDOs and more related to an FDO ecosystem providing such registries. What RO-Crates would need to do wrt **FDOR7** is making sure to keep up-to-date with the pairs type-operations maintained in such registries. For this, an external agent, e.g., RO-Crate monitor, would be needed.

RO-Crates support packaging metadata so it can be an FDO itself, **FDOR8**. RO-Crates allows metadata of different types, **FDOR9**, as far as they follow an existing profile. **FDOR10**, metadata schemas are FDOs maintained by communities, needs some additional work as Bioschemas specifications are not currently compliant to FDOs. Collections are supported by RO-Crates, **FDOR11**, but their construction is not yet part of the RO-crates. Finally, deletion mechanisms, **FDOR12**, are not yet supported by RO-Crates.

3. Possible Use and Benefits in NFDI

Being FAIR a cornerstone for NFDI consortia, FDOs would be an enhancement as the operational level should improve (semi)automatic interoperability across multiple disciplines and sorts of DOs. FDO implementation via RO-Crates provides an already known environment as it relies on LD, schema.org and Bioschemas, with some NFDIs already using schema.org (e.g., NFDI4DataScience, NFDI4Chem, NFDI4Culture, NFDI4MatWerk). As RO-Crates rely on LD, they are also compatible with Knowledge Graphs, a fundamental element in many NFDIs. In addition, RO-Crates flexibility, particularly regarding the metadata accompanying the actual DO, becomes an advantage for a multidisciplinary environment such as the NFDI consortia. RO-Crates can also be combined with authorization and authentication so private data remains protected (e.g., by pointing to a protected repository rather than including the data) while its metadata is open and public.

3. Future work

RO-Crate is a promising route towards FDOs. Although it does not yet fulfill all the FDORs, there is a growing interest in the RO-Crate community in supporting the realization of FDOs and benefit of their advantages with some ongoing efforts to that end [8].

Data availability statement

This submission is not based on data.

Author contributions

LJC: conceptualization, writing – original draft, writing – review & editing. SSR: conceptualization, project administration, writing – review & editing. DRS: conceptualization, funding acquisition, project administration, writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

LJC and DRS are supported by the NFDI4DataScience, funded by the Deutsche Forschungsgemeinschaft DFG, project no. 460234259. SSR is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement numbers 823830 (BioExcel-2), 824087 (EOSC-Life) and the Horizon Europe programme under grant agreement 101046203 (BY-COVID).

References

1. C. Bizer, T. Heath, and T. Berners-Lee, 'Linked Data - The Story So Far', International Journal on Semantic Web and Information Systems (IJSWIS), vol. 5, no. 3, pp. 1–22, 2009, doi:<https://doi.org/10.4018/jswis.2009081901>
2. A. Hasnain and D. Rebholz-Schuhmann, 'Assessing FAIR Data Principles Against the 5-Star Open Data Principles', in The Semantic Web: ESWC 2018 Satellite Events, A. Gangemi, A. L. Gentile, A. G. Nuzzolese, S. Rudolph, M. Maleshkova, H. Paulheim, J. Z. Pan, and M. Alam, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 469–477. doi:https://doi.org/10.1007/978-3-319-98192-5_60
3. M. D. Wilkinson et al., 'The FAIR Guiding Principles for scientific data management and stewardship', Sci Data, vol. 3, no. 1, Art. no. 1, Dec. 2016, doi:<https://doi.org/10.1038/sdata.2016.18>
4. K. De Smedt, D. Koureas, and P. Wittenburg, 'FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units', Publications, vol. 8, no. 2, Art. no. 2, Jun. 2020, doi:<https://10.3390/publications8020021>
5. S. Soiland-Reyes et al., 'Packaging research artefacts with RO-Crate', Data Science, pp. 1–42, Jan. 2022, doi:<https://doi.org/10.3233/DS-210053>
6. R. V. Guha, D. Brickley, and S. Macbeth, 'Schema.org: evolution of structured data on the web', Commun. ACM, vol. 59, no. 2, pp. 44–51, Jan. 2016, doi:<https://doi.org/10.1145/2844544>.

7. S. Soiland-Reyes, L. J. Castro, D. Garijo, M. Portier, C. Goble, and P. Groth, 'Updating Linked Data practices for FAIR Digital Object principles', in Research Ideas and Outcomes, Pensoft Publishers, Oct. 2022, p. e94501. doi:<https://doi.org/10.3897/rio.8.e94501>
8. S. Soiland-Reyes et al., 'Creating lightweight FAIR Digital Objects with RO-Crate', in Research Ideas and Outcomes, Pensoft Publishers, Oct. 2022, p. e93937. doi:<https://doi.org/10.3897/rio.8.e93937>
9. G. Strawn, P. Wittenburg Eds. "FDO Forum FDO Requirement Specifications Version 1.0". Available at https://docs.google.com/document/d/1-4_yGRrlcgdMIwaFvHyUt6lxDdfzGpqLUCehE0vJ-q (accessed 2023.04.26)
10. DOI Foundation. "DOI Foundation". <https://www.doi.org/> (accessed 2023.04.26)
11. W3ID Community. "Permanent Identifiers for the Web". <https://w3id.org/> (accessed 2023.04.26)
12. A. J. G. Gray, C. Goble, and R. C. Jimenez, 'From Potato Salad to Protein Annotation', in ISWC Posters and Demo session, Vienna, Austria, Oct. 2017, p. 4. [Online]. Available: <http://ceur-ws.org/Vol-1963/paper579.pdf>

ICT Infrastructure supporting the Italian Research Infrastructure on Microbial Resources MIRRI-IT

Marco Beccuti¹[\[https://orcid.org/0000-0001-6125-9460\]](https://orcid.org/0000-0001-6125-9460), Antonio d’Acerno²[\[https://orcid.org/0000-0003-0516-0794\]](https://orcid.org/0000-0003-0516-0794),
Simone Donetti¹[\[https://orcid.org/0000-0001-7204-4078\]](https://orcid.org/0000-0001-7204-4078), Sandro Gepiro Contaldo¹[\[https://orcid.org/0009-0007-3889-8644\]](https://orcid.org/0009-0007-3889-8644), Paolo Romano³[\[https://orcid.org/0000-0003-4694-3883\]](https://orcid.org/0000-0003-4694-3883), and Giovanna Cristina Varese¹[\[https://orcid.org/0000-0002-1455-6208\]](https://orcid.org/0000-0002-1455-6208)

¹ University of Turin, IT

² ISA - CNR, Italy

³ IRCCS Ospedale Policlinico San Martino, Genoa, Italy

Keywords: Research Infrastructure, Microbial resources.

Extended Abstract

In 2022, the SUS-MIRRI.IT project (Strengthening the MIRRI Italian Research Infrastructure for Sustainable Bioscience and Bioeconomy, www.sus-mirri.it) was funded with ~17M€ by the Italian government on the NextGeneration EU-funded Recovery and Resilience National Plan (PNRR) – Research Infrastructure - to strengthen the Italian Research Infrastructure for Microbial Resources (MIRRI-IT) and ensure its long-term sustainability.

The main objectives of this project are:

- (a) to implement MIRRI-IT’s organisation and set up its operative procedures and quality standards;
- (b) to improve quality of Italian microbial Biological Resource Centers databases and conceive MIRRI-IT’s services based on the partners’ expertise and genetic resources;
- (c) to set up a single entry point platform to promote MIRRI-IT’s resources in terms of data, services, cutting-edge technologies and expertise.

In this contest, the new needs for data integration and system interoperability, stressed by the FAIR approach to data sharing, have become evident and urgent thus leading to the identification of four ICT macro-activities which will be carried out in the project.

- (1) The activity related to the Italian Collaborative Working Environment (ItCWE) platform foresees its implementation based on WordPress [1], a very popular content management system. It will provide four main functionalities, namely access to (i) data and services, (ii) expert consulting, (iii) TransNational Access (TNA) program, and (iv) training. The WordPress multisite instance will be configured taking as a reference the platform recently implemented for the Microbial Resource Research Infrastructure MIRRI.
- (2) The activity related to the Italian Culture Collections Catalogue (ItCCC) aims to develop it as the main access point to information on microbial resources available at the national level. As such, it must be as FAIR (Findable, Accessible, Interoperable, Reusable) as possible, while taking into account the existing limitations of microbial information semantics. The ItCCC will consider the data model agreed upon by the ICT

Task Force of MIRRI at the European level and defined in the context of the IS_MIRRI21 EU project (<https://ismirri21.mirri.org/>) as the reference dataset for the MIRRI Information System (MIRRI-IS), to facilitate the upload of data from Italian collections into the ItCCC and, possibly, the MIRRI-IS. Data will be also made available to researchers through dynamic web pages, while interoperability will be ensured using standard technologies (such as REST APIs, JSON). ItCCC will be served by the Apache HTTP server [2] and based on PostgreSQL [3] and PHP scripts [4].

- (3) The activity related to the Microbial Biological Resource Center database (mBRCdb) will lead to the development of a standalone open-source application supporting the local management of their catalogue by single culture collections (CCs). Indeed, many CCs do not currently have an effective data management system for their collection and do not have IT staff able to support its implementation. This application will also guide the curators transforming data into a format compliant with the MIRRI-IS data model by implementing basic data quality controls. To maximize its portability on different operating systems and architectures, Java language was identified as an appropriate candidate for its development.
- (4) The Dataverse proof-of-concept (PoC) activity is related to the implementation of a Harvard Dataverse [5] instance to investigate how the semantic annotation of microbial information could be suitably stored by exploiting this open-source tool that is reputed as one of the most effective tools for improving data FAIRness. Its configurable system of roles and nested repositories will be investigated to see if it could allow the enforcement of data curation, preservation, and publishing workflows. Moreover, its ability to integrate with established user identification platforms, such as ORCID, and digital objects, such as DOI and FDO, will be investigated. Its capacity to share and discover data through semantic metadata will also be investigated..

The corresponding computational infrastructure that we implemented for hosting these ICT activities is reported in Fig.1 where the hardware level is highlighted in gray, the virtualization level in dark orange, and the application level corresponding to the macro-services in blue and the micro-services needed for each macro-services in a light orange.

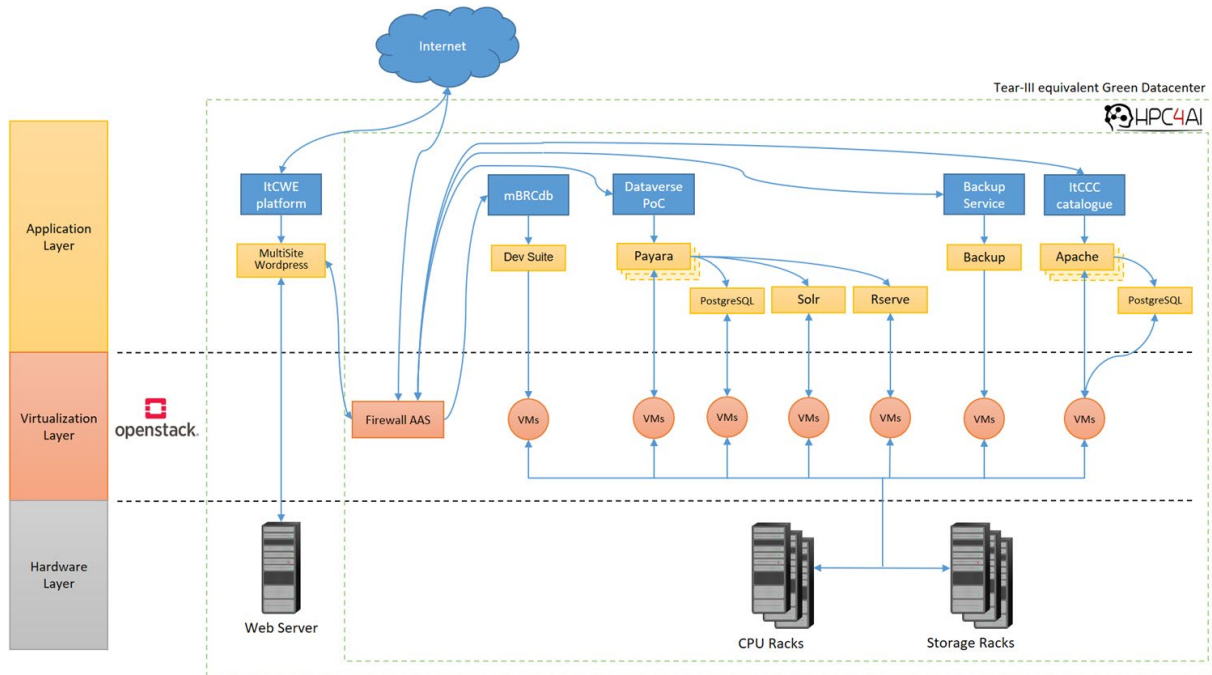


Fig.1: The schema of ICT infrastructure implemented for the SUS-MIRRI.IT project.

The Hardware level was set up on the HPC4AI data centre (<https://hpc4ai.unito.it>) based on HPC-cloud convergence architecture with 2200+ cores, 100+ GPUs, and four storage classes (i.e. 300TB SSD hyper-converged, 1PB hybrid multi-tenant HW-encrypted scale-out NAS, 1PB cold storage backup, 170TB deduplicated storage backup), 100GB/s SDN networking. It is organized in experimental and production islands hosted into 250KVA Tier-III adiabatic green datacenter with 16 racks.

The virtualization level is based on the open-source OpenStack cloud technology (co-developed with Canonical - <https://canonical.com/>, the company behind Ubuntu), which is currently the de facto standard for HPC-cloud convergence architecture. Indeed, OpenStack manages data centre resources through a set of modular services which interoperate in a service-oriented architecture. It provides abstractions for computational, storage, and networking resources and allows cloud administrators to provide them easily, dynamically, and securely to different tenants.

The application level implements the project macro-services in terms of their microservices exploiting the OpenStack modularity which allows us to easily add new services or rescale those already created according to new requirements that would arise during the project.

Initially all the VMs hosting the microservices were created with a medium profile of resource assignment (i.e. 4 vCPU, 8GB RAM), while the disk space allocation is managed dynamically according to the amount of data that will have to be stored and managed. All disk space used is provided by the cloud as virtual space, with redundant management of physical disks using different technologies. Depending on the type of disk chosen, the management policy changes and we pass from Ceph-based software systems with Replica 3 or Erasure Code¹ ($k=2, m=1$) to dedicated hardware appliance solutions with mixed-type disk storage or appliance with hardware deduplication. VMs use disks of all types depending on the needs and tasks to be performed.

Author contributions

All authors contributed equally

Competing interests

The authors declare that they have no competing interests.

Funding

SUS-MIRRI.it PNRR project.

References

1. WordPress Official Site. URL: <https://wordpress.com/>
2. Apache HTTP Server Project Official Site. URL: <https://httpd.apache.org/>
3. PostgreSQL Official Site. URL: <https://www.postgresql.org/>
4. PHP Official Site URL: <https://www.php.net/>

¹ Erasure coding uses storage capacity more efficiently than replication. The n-replication approach maintains n copies of an object (3x by default in Ceph), whereas erasure coding maintains only $k + m$ chunks. For example, 2 data and 1 coding chunks use 1.5x the storage space of the original object.

5. Gary King, "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing", *Sociological Methods & Research*, vol. 36(2), pages 173-199, November 2007.

Enhancing Reproducibility in Research through FAIR Digital Objects

Zeyd Boukhers¹[\[https://orcid.org/0000-0001-9778-9164\]](https://orcid.org/0000-0001-9778-9164), and Leyla Jael Castro³[\[https://orcid.org/0000-0003-3986-0510\]](https://orcid.org/0000-0003-3986-0510)

¹Fraunhofer Institute for Applied Information Technology, Sankt Augustin, Germany

²University Hospital of Cologne, Cologne, Germany

³ZB MED Information Centre for Life Sciences, Cologne, Germany

Abstract:

The FAIR principles were introduced to enhance data reuse by providing guidelines for effective data management practices. In the broader context of research, assets encompass not only data but also other artifacts such as code, software, and publications. FAIRifying these artifacts is as essential as FAIRifying data, especially in Data Science and Artificial Intelligence, where the complexity of current AI approaches makes reproducibility extremely challenging. Therefore, facilitating the easy reuse of these artifacts represents a significant stride towards mitigating this problem. The concept of FAIR Digital Objects (FDOs) presents a solution to FAIRify these artifacts, treating them as FDOs. NFDI4DataScience is embracing FDOs and proposing an architecture to efficiently manage them.

Keywords: FAIR Digital Object, Reproducibility, Research Data, Metadata, NFDI4DS

1 Background

Since introducing the FAIR principles in 2016 [1], they have gained widespread acceptance among data management professionals, projects, and initiatives, including the EOSC roadmap (European Commission 2018). These principles were later expanded to encompass other digital objects (e.g., software [2] and workflows [3]). Recognizing the importance of these principles such as making metadata available, significant efforts have been mobilized to motivate researchers to FAIRify their resources and supply metadata, with various strategies employed to extract metadata from existing resources such as publications [4]. However, it's become clear that availability alone is insufficient to achieve the desired level of reproducibility and reusability. Metadata must not only be accessible but also complete, accurate, and readily findable. The FAIR Digital Objects (FDOs) emerged as a more detailed version, where objects, their metadata and their materialisations are clearly separated. An FDO is a conceptual approach [5] that offers a technical solution for implementing FAIR principles across various digital

object types [6], [7] while also adding an operations layer boosting interoperability beyond the initial FAIR possibilities as machines can understand and directly use those operations to (semi)automatically use the objects and connect them with each other via operational workflows.

Numerous initiatives have emerged in Europe and beyond to establish a comprehensive infrastructure ecosystem for research data, such as the EOSC at the European level and NFDI at the German national level. It is crucial that these infrastructures are founded on robust fundamentals, like those offered by FDOs. The implementation of FDOs helps ensure that research digital objects are findable, accessible, interoperable, and reusable, thus paving the way towards improved transparency, and reproducibility in data-driven research disciplines.

2 FDO Concept

The adoption of FAIR Digital Objects (FDOs) enables researchers and institutions to effectively manage digital objects while ensuring compliance with evolving research management standards. This, in turn, enables more effective sharing and collaboration among researchers, fostering innovation and accelerating scientific discovery. Additionally, as more funding agencies and publishers begin to require adherence to FAIR principles, the adoption of FDOs will become increasingly important to secure funding and disseminate research findings. Consequently, the implementation of FDOs can create a foundation for more transparent, efficient, and collaborative research processes that ultimately benefit the entire scientific community.

Recognizing the importance of FDOs, the NFDI4DataScience¹ consortium is adopting them to manage research artefacts within the realm of data science, including essential components such as datasets, publications, software, and models. The objective is to assist researchers in easily locating and utilising the resources they need, whether they correspond to the metadata or the actual object in a particular format together with some basic operations on top of them. This is particularly useful when the objects are private as their metadata will still be open and available to others.

3 FDO Architecture

In accordance with the FDO specifications² released by the FDO Forum³, we propose an architecture⁴ that adheres to both FAIR principles and FDO specifications while encompassing all artifacts in a data science workflow. As illustrated in Figure 1, the architecture begins with a Registry of FDOs that stores all the PID records along with minimal metadata records such as the FDO type.

Each PID in the registry resolves to a more extensive metadata piece containing the FDO Profile, the PID of the metadata, and the PID of the digital object itself (e.g., publication, dataset). The rationale for having separate PIDs for the digital object and metadata is to comply with the FDO specification, which emphasizes separating the metadata from the digital object and ensuring that metadata persists even if the digital object does not.

¹<https://www.nfdi4datascience.de/>

²<https://fairdo.org/specifications/>

³<https://fairdo.org/>

⁴<https://fdda1.gitlab.io/fdom>

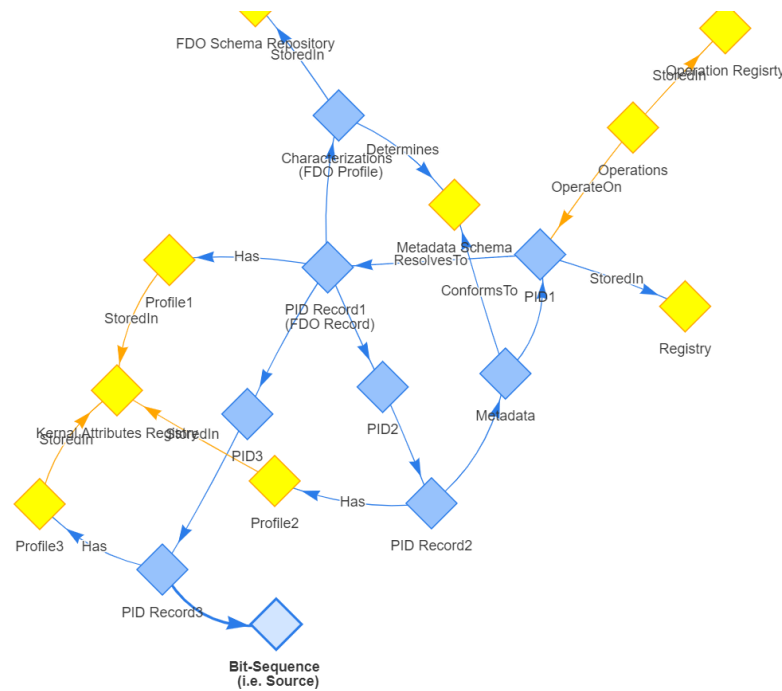


Figure 1. FDO Architecture. See footnote 4

While the digital object PID resolves to the bit-sequence, the metadata PID resolves to the corresponding metadata of the FDO, which follows predefined metadata schemas. These schemas are determined by the FDO profile, such as software or dataset, and obtained from Schema.org⁵. Additionally, a set of operations from a predefined registry is assigned to each FDO based on its profile.

This architecture fosters compliance with FAIR principles and FDO specifications, ensuring that all artifacts in a data science workflow are managed effectively. An example of this architecture can be found here⁶.

Author contributions

Z.B: conceptualization, writing – original draft, writing – review & editing. LJC: conceptualization, project administration, writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

This work received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number: NFDI4DataScienc (460234259).

⁵<https://schema.org/>

⁶https://ai-research.net/FDO_Example.html

References

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [2] M. Barker, N. P. Chue Hong, D. S. Katz, *et al.*, "Introducing the fair principles for research software," *Scientific Data*, vol. 9, no. 1, p. 622, 2022.
- [3] C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, *et al.*, "Fair computational workflows," *Data Intelligence*, vol. 2, no. 1-2, pp. 108–121, 2020.
- [4] Z. Boukhers, N. Beili, T. Hartmann, P. Goswami, and M. A. Zafar, "Mexpub: Deep transfer learning for metadata extraction from german publications," in *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDDL)*, IEEE, 2021, pp. 250–253.
- [5] K. De Smedt, D. Koureas, and P. Wittenburg, "Fair digital objects for science: From data pieces to actionable knowledge units," *Publications*, vol. 8, no. 2, p. 21, 2020.
- [6] E. Schultes and P. Wittenburg, "Fair principles and digital objects: Accelerating convergence on a data infrastructure," in *Data Analytics and Management in Data Intensive Domains: 20th International Conference, DAMDID/RCDL 2018, Moscow, Russia, October 9–12, 2018, Revised Selected Papers 20*, Springer, 2019, pp. 3–16.
- [7] U. Schwardmann, "Digital objects—fair digital objects: Which services are required?" *Data Science Journal*, vol. 19, no. 1, 2020.

Automated Documentation of Research Processes Using RDM

Lars C. Griem^{1,2}, Richard Thelen³, and Michael Selzer²

¹ Institute for Applied Materials – Microstructure Modelling and Simulation (IAM-MMS), Karlsruhe Institute of Technology (KIT), Germany

² Institute of Nanotechnology – Microstructure Simulation (INT-MSS), Karlsruhe Institute of Technology (KIT), Germany

³ Institute of Microstructure Technology (IMT), Karlsruhe Institute of Technology (KIT), Germany

Abstract. Published research results usually represent only a fraction of the data generated at a research institute. The unpublished data created in the process of producing the final result, however, often contain valuable information that can be reused. Through research data management, all these data should be stored centrally according to the FAIR principles (Findable, Accessible, Interoperable, Reusable). However, a significant part of knowledge is often not found in the data, but in the processes that led to their generation. It is therefore important to map these processes to archive and document this knowledge in a structured way. Procedures for documenting scientific processes already exist and are actively used at research institutes. However, these are often analogue or paper-based and hence do not meet the requirements for FAIR data management. At the Institute for Microstructure Technology of the KIT, such a paper-based procedure is used to document the production of microstructure components. During their manufacturing, it is essential to adhere to the correct process parameters in order to enable error-free production. Therefore, a so-called job ticket always accompanies the production of components. On this job ticket, the correct process sequence is listed and a detailed description of the respective process step is given. Depending on the component to be produced, a distinction is made between different types of job tickets according to internal conventions. On the one hand, there are so-called green job tickets, which describe a standardised process sequence, and on the other hand, blue job tickets, which are intended to document experimental manufacturing processes. The process sequence on the blue job tickets is initially empty and is filled in during the manufacturing process. Common to both types of job tickets is that they are stored in the institute's archive after completion of the component production. However, since the job tickets are paper-based, the corresponding archive of job tickets cannot be searched quickly and, given the sheer volume of archived job tickets, represents an unmanageable collection of data. The existing system for process documentation is therefore to be implemented with the help of the research data infrastructure Kadi4Mat [1] in accordance with FAIR principles, thereby making the available process knowledge more accessible.

1. Methods

The Kadi4Mat research data infrastructure provides a set of software modules that address different aspects of research data management. These include KadiWeb, which provides a repository for structured data storage, and KadiStudio, a workflow editor that enables reproducible documentation of research processes in the form of automated workflows.

In KadiWeb, research data is always stored in the form of records. These are containers in which research data is stored and provided with descriptive metadata. The design of the metadata is left to the users and can therefore be adapted to all possible use cases and metadata schemes. In order to ensure the FAIR storage of research data, it is necessary that uniform metadata schemas are used within research institutions. For this purpose, metadata templates exist that can be used when creating records and assigning defined metadata to them.

The workflow editor KadiStudio is based on the concept that every scientific process can be divided into a multitude of atomistic, not further divisible, simple work steps [2]. These steps are available in KadiStudio as executable functions and can be freely recombined to model a workflow using an intuitive graphical user interface. The library of available nodes can be extended by the user as desired, allowing KadiStudio to be continuously developed and adapted to specific use cases.

2. Results

The two modules described above, KadiWeb and KadiStudio, are used to model the job ticket documentation. KadiWeb is used to store and archive the job tickets in a machine-readable format, while KadiStudio is used to actually fill in the job ticket with work instructions.

In order to achieve uniform documentation of the job tickets, templates were first created on KadiWeb. Depending on the type of job ticket, green or blue, different approaches were taken. For green job tickets, templates were created in which all the individual steps of the manufacturing process were defined in the metadata as shown in Figure 1.

Typ	Value	Type	Action
Nummer der Laufkarte	null	Integer	✎
Losnummer	null	String	✎
IFA-Nummer	null	String	✎
Erzeugnisnummer	null	String	✎
Auftraggeber	null	String	✎
Beistellteile, Masken etc.	null	String	✎
Fertigungsgegenstand	null	String	✎

Figure 1. Green Job Ticket Template in KadiWeb.

The blue job ticket template, only contains basic metadata and no processing steps. Instead, additional templates were created for all possible processing steps. These can be added to a blue job ticket to document the manufacturing process. All metadata entries were further

equipped with sanity checks to ensure that only meaningful data can be entered. To automate the documentation of job tickets using these templates, a workflow in KadiStudio was used, shown in Figure 2.

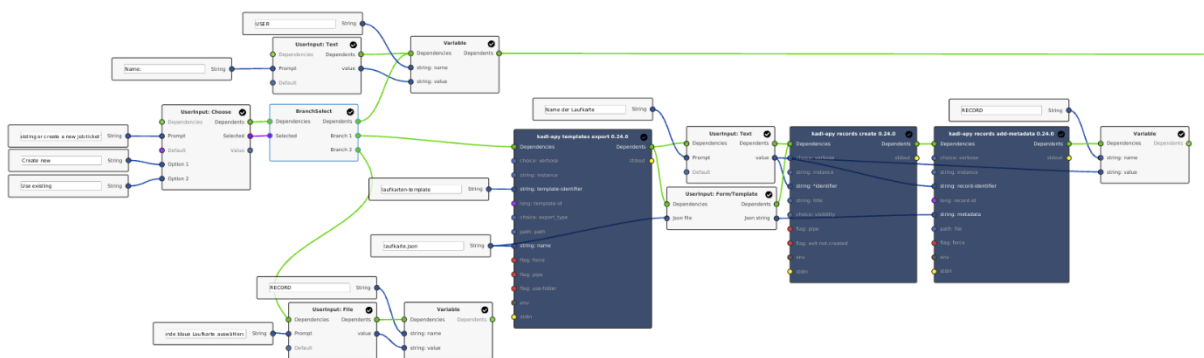


Figure 2. Created job ticket workflow in KadiStudio.

When executing this workflow, the user is first asked whether a green or blue job ticket is to be processed, and is then given the option of creating a new job ticket record in KadiWeb or continuing to work on an existing record. To do this, KadiStudio communicates with KadiWeb in the background and loads all available job ticket records and templates for the user to choose from. The selected job ticket is then used for further processing. In the case of a green job ticket, the next processing step is automatically retrieved from the template. For blue job tickets, the user can interactively select the next step from all available working step templates. In both cases, KadiStudio creates an input mask shown in Figure 3 from the corresponding template and prompts the user to fill it in. Once the mask has been filled in, the metadata is added to the record of the job card in KadiWeb. In this way, the documentation of job tickets is completely digitalised and automated.

Figure 3. User prompt generated from the metadata templates.

3. Conclusion

Integrating the job ticket documentation into the research data infrastructure Kadi4Mat offers many benefits. The job tickets can be quickly searched for specific parameters, equipment or processing. In addition, the user is interactively guided through the manufacturing process in the workflows. This makes it easier for new employees to get started and ensures quality assurance with the help of sanity checks.

Data availability statement

-

Underlying and related material

-

Author contributions

-

Competing interests

The authors declare no competing interests.

Funding

-

Acknowledgement

-

References

1. Brandt, Nico et al. Data Science Journal, vol. 20, 2021, p. 8.
2. Griem, L, et al. Data Science Journal, vol. 21, 2022, p. 16.

Improving the research desktop experience for OpenStack VDI

Integrating hardware accelerated rendering and remote transport

Dirk von Suchodoletz¹[\[https://orcid.org/0000-0002-4382-5104\]](https://orcid.org/0000-0002-4382-5104), Yi Sun²[\[https://orcid.org/0000-0002-7636-0200\]](https://orcid.org/0000-0002-7636-0200),
Jean-Karim Hériché²[\[https://orcid.org/0000-0001-6867-9425\]](https://orcid.org/0000-0001-6867-9425), and
Manuel J. Messner¹

¹Computer Center, University of Freiburg, Germany

²EMBL, Heidelberg, Germany

Abstract: Using virtual machines with dedicated rendering and remote access capabilities, virtual workplaces for various usecases can be created, and then accessed from anywhere at any time. If this is to happen on a large scale in the cloud, so-called VDI for the dynamic provision of virtual desktops play an increasingly important role. A sustainable VDI should be freely available to everyone for modification and redistribution at no cost, be scalable, and should support various desktop use cases with different resource requirements. Some use cases involve hundreds of similar VM running in parallel, which requires proper resource planning.

Keywords: Virtual Desktop Infrastructure, hardware accelerated rendering, remote access, NFDI4BioImage, DataPLANT

1 Motivation

The objective of this initiative in the **Enabling RDM track** is primarily to form a special interest group for hardware accelerated VDI on cloud infrastructure as we expect the uptake of the topic in other consortia as well. For various use cases in research, education and training Virtual Desktop Infrastructure (VDI) would provide an efficient means in the context of the NFDI. The available commercial frameworks are quite expensive (for large and scattered audiences) and do not necessarily cater to the core needs of researchers. Thus we are looking for Open Source alternatives based on well established software frameworks. We identified three relevant use cases:

1. Provide remote access to desktop environments to handle large scale high resolution imaging data in various research domains (near to the location of the data to avoid tedious transfers) and Remote visualization in High Performance Computing
2. Streamline training resources to allow more flexible remote teaching and working models for university staff in research
3. Provide controlled access to sensitive data in a protected environment

Many computer centers already offer OpenStack cloud infrastructure for various purposes. In Freiburg we are investigating the options to transfer our experience from our decentralized bwLehrpool VDI solution to OpenStack to offer an Open Source backed VDI (OSVDI).

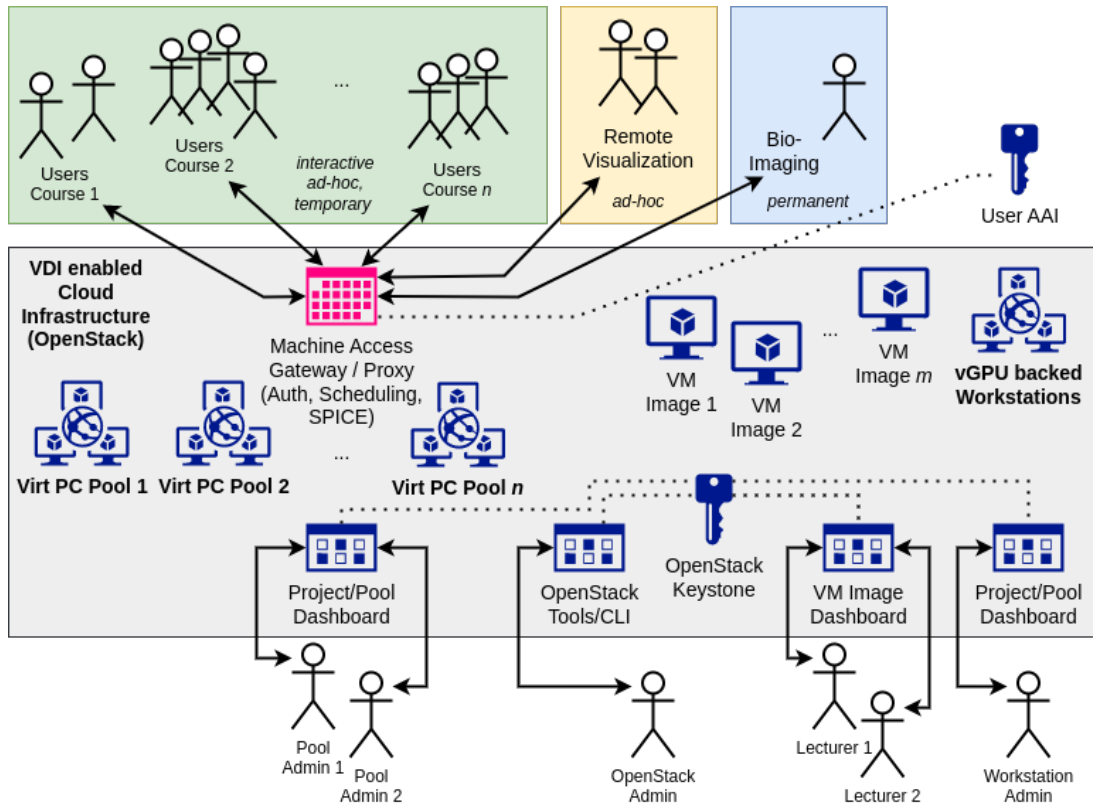


Figure 1. Various use cases and general infrastructure framework of the OSVDI framework.

OpenStack is a free cloud platform, most commonly deployed as Infrastructure-as-a-Service. The platform is composed of several components and services. Each component and service is responsible for a certain set of tasks and provides a RESTful API for communication. OpenStack’s main focus is to provide an infrastructure for VMs, their storage and their network. Besides the mediated GPU passthrough support in OpenStack we need resource scheduling for cloud computing.

2 Objective

Our vision for a next generation approach is to have a public VDI web application where users log in, select a VM they want to use, and finally be scheduled to an appropriate cloud node. Additional logic could be added, e.g. limiting or skipping the VM selection for a researcher or training group depending on their location or time of day. Our focus is mainly on the mediated GPU passthrough for further development of an OSVDI because this approach combines the flexibility of emulation and paravirtualization with the performance boost of direct GPU passthrough. Linux as the host system already provides the *mdev* subsystem and tools for mediated devices (vGPUs) and their device drivers. Using this subsystem has the major advantage that the Linux host system can manage all vGPUs and mediate shared access. The direct access to a framebuffer of a vGPU means in terms of an OSVDI that the Linux host system can obtain the graphics output of any VM (virtual desktop) and can control those output for further processing and transfer to remote (thin) clients. Access to a VM session can then be

implemented via two methods: A browser-based approach, using modern technologies like WebAssembly, WebUSB and MediaDevices, resulting in immediate access from a wide range of devices. Still, the alternative approach of using a dedicated native application the user has to install first can offer even greater integration with the user's system, as well as yielding better performance depending on its use case.

Our goal is to structure the development of the missing bits and pieces which got identified by a paper published in 2021 [1]:

- Defining and implementing an access gateway to VDI on OpenStack clouds
- Evaluate and improve the existing remote access protocol SPICE and the associated web and native clients
- Adapt OpenStack scheduling to allow the hosting of the relevant usecases and in-advance reservation of resources for e.g. trainings [2]
- Coordinate these efforts with other projects especially in the context of EOSC interested in this domain

3 Envisioned milestones

For the imagined OSVDI we plan several development cycles and a minimum viable product approach. In a precursor the existing Guacamole bwLehrpool remote access should get improved through hardware rendering and stream encoding deploying the Intel GVT-g desktop graphic architecture together with the KVM infrastructure as a Linux-based hypervisor and produce an assessment of ease to use and stability. This will get implemented as an enhanced bwLehrpool service and demonstrate the capabilities of the existing kernel drivers regarding GPU virtualization and hardware partitioning. We will use the SPICE client and Looking Glass as a test and performance measure when accessing the virtual framebuffer for AVC/H.264 encoding and transport. Upon this we will explore how to encode with low latency, and how to send it to browsers and display the content there with low latency. This provides a possible baseline to check certain expectations and features before delivering similar services like those for an OpenStack cloud.

In a further milestone, we focus on a basic VNC model (leaving further improvements of remote access to parallel or later developments) in the cloud including orchestration of resources which covers the scope of our contribution to NFDI4BIOIMAGE. This milestone starts to extend the OpenStack framework for missing components and modules. This milestone deals with the challenges of a suitable access broker to distribute users requesting certain types of desktops onto a suitable VM. The access broker includes the provisioning of basic interaction channels starting from the input of the various usecases. While the previous step focused on a basic integration and the outline of strategic components the next milestone focuses on the special hardware virtualization and integration parts both from the viewpoint of the guest systems and as encoding devices from the host perspective. The remote access should enjoy at least an enhanced hardware-backed video stream transport model for the remote visual cloud. Later milestones should deal with further remote interaction channels and further features and improvements for typical VDI setups like suspend and resume of interactive desktop sessions. Starting during the second milestone measures should be taken to form a sustainable community and financing concept around the proposed service. Both ongoing support, code maintenance and future development are to be supported through some stable organizational structure.

Competing interests

The authors declare that they have no competing interests.

Funding

is funded by the German Research Foundation (DFG) within the framework of the NFDI – project number: 442077441 and EOSC-Life under grant agreement H2020-EU.1.4.1.1. EOSC-Life 824087.

Acknowledgements

Part of the activities and insights presented in this paper were made possible through the collaboration in the PePP project (FBM2020-VA77-8-01241) funded by the *Foundation for Innovation in Higher Education*, EOSC-Life and preliminary work in the bwLehrpool project supported by the Baden-Württemberg Ministry of Science, Research, and the Arts.

References

References

- [1] M. Bentele, D. von Suchodoletz, M. Messner, and S. Rettberg, "Towards a GPU-Accelerated open source VDI for OpenStack," in *Cloud Computing*, M. R. Khosravi, Q. He, and H. Dai, Eds., Cham: Springer International Publishing, 2022, pp. 149–164, ISBN: 978-3-030-99191-3. DOI: [10.1007/978-3-030-99191-3_12](https://doi.org/10.1007/978-3-030-99191-3_12).
- [2] M. Bentele, D. von Suchodoletz, M. Messner, and R. Piliszek, *A resource-aware scheduling concept for an openstack-based vdi*, Accepted for publication at CloudComp 2023 - 12th EAI International Conference on Cloud Computing, 2023.

DataPLANT Cloud Oriented Service Infrastructure

Open for Integration and Adaptation

Dirk von Suchodoletz¹[\[https://orcid.org/0000-0002-4382-5104\]](https://orcid.org/0000-0002-4382-5104), Jonathan Bauer²[\[https://orcid.org/0000-0002-5624-2055\]](https://orcid.org/0000-0002-5624-2055),
and Marcel Tschöpe³[\[https://orcid.org/0000-0002-3731-7664\]](https://orcid.org/0000-0002-3731-7664)

¹ Computer Center, University of Freiburg, Germany

Abstract. A core objective of the DataPLANT consortium is to provide a science gateway as a set of flexible cloud-based (micro) services. The setup aims at both on-premises installations and future integration into a shared NFDI infrastructure. We will present the DataPLANT DataHUB, which provides various RDM workflows to support research data scientists at different stages of the data lifecycle - from development to publication of the results obtained.

Keywords: FAIR Data sharing, Service Infrastructure, RDM Platform, cloud deployment, GitLab, InvenioRDM

1. Introduction

The aim of this presentation in the **Enabling RDM** track is to exchange ideas with other NFDI consortia on the services required for RDM and the principles for service development and deployment. These considerations can be used as input for joint infrastructure development, e.g. in the context of the NFDI Common Infrastructure section or Base4NFDI. Over the past two and a half years, the DataPLANT team has developed a set of software and system components that provide services to the basic plant research community [1,2]. The set of tools and microservices that have been developed and evolved to date have focused on extending the existing digital landscape of the typical plant scientist. The core services focus on data management, versioning, sharing and publishing. All services are centered around cloud deployable modules.

The development of applications and tools to support community-driven research data management requires the involvement of several parties. During the development of the services, we agreed on design principles to provide a high-level guidance and a set of criteria for creating desirable and maintainable applications. In DataPLANT, tool development is always motivated by community requirements, conveyed by researchers, e.g., through data stewards, to developers [3]. Developments in DataPLANT follow an incremental and iterative approach, ensuring commitment and alignment of expectations of all stakeholders. Another aim of the service development is to enable both central and local installation of services without divergent implementations. Thus, we hope to encourage adoption by other communities and integration into a future NFDI service infrastructure.

2. Basic Infrastructure

The basic infrastructure is the "Persistent/Dynamic" layer of Figure 1, which consists of the required storage and compute resources, which are provided by the de.NBI cloud in our setting. Optimally, the hosting infrastructure for virtual machines (VMs) or containers allows for automation, e.g., to redeploy a service after a failure or update.

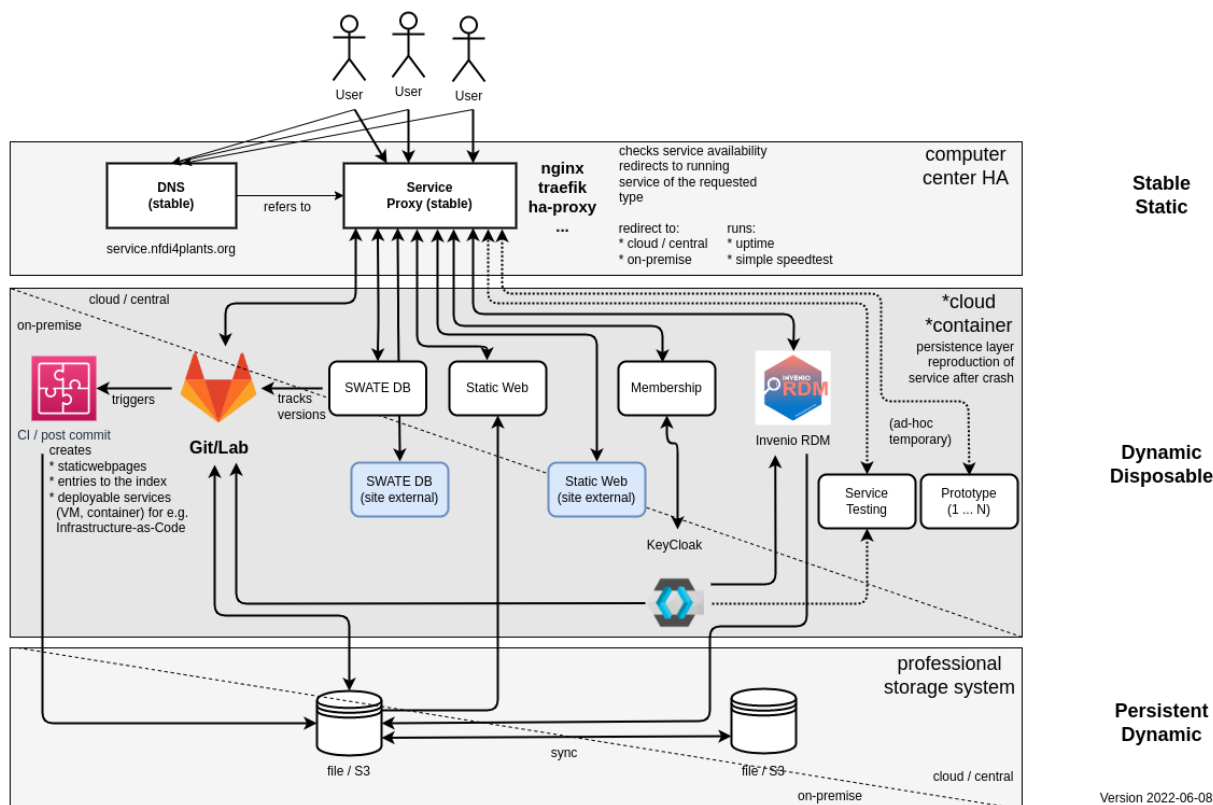


Figure 1. The DataPLANT services infrastructure building blocks.

Storage resources are used to store both the necessary service configurations and user data. Depending on the service provided locally, storage resources may be provided in the form of traditional network file systems such as NFS or SMB, or as object storage. The storage service implements the necessary redundancy to provide the required level of security for user data.

An authentication instance is required to authenticate users and the services behind them. The DataPLANT user management builds on existing AAIs. Well established services such as the Life Sciences AAI and ORCID can be combined with local authentication within the central DataPLANT authentication service. The infrastructure is based on KeyCloak, which supports modern authentication protocols such as OpenID Connect and SAML, enabling the integration of multiple AAIs and identity brokering. Providing AAI identity management that can easily connect to GitLab and other services using either protocol simplifies user management. Connecting multiple AAIs through KeyCloak allows our community to use their existing accounts, such as the Life Sciences AAI, their home institution or ORCID. We can assign different roles depending on the source of the account or specific attributes. Permissions can be derived from these roles to differentiate between users. These range from privileged users with full access to the data and the ability to create archives/publications, to users who only have a reporting function and/or read-only access to raw data.

The DataPLANT microservices are packaged as Docker containers running in the "dynamic/disposable" layer in Figure 1. This allows for a flexible deployment strategy, as the containers can be run both on VMs and/or directly in a container runtime environment such as Kubernetes or OpenShift. Depending on the local requirements for on-premises services, only a subset of the DataPLANT services will be considered.

The core helper services in DataPLANT consist of Traefik as a proxy service, Keycloak as an identity and access management framework for user management, and Kuma as a service monitoring framework. The entry proxy for DataPLANT services performs a number of functions, providing a stable point for users to connect to regardless of where individual services are actually geographically located and running. Another intended feature is that the proxy has the ability to trigger the (re)deployment (to a suitable infrastructure) if it is not available. The proxy can be used to redirect to other network destinations (outside the site itself) depending on the source institution (so you can transparently implement partially on premise). It is the established connection point that makes it easy to test new services. For these tasks it would be desirable to be able to configure the proxy on-the-fly via an API (should have a simple reload mechanism). Host helper services such as speedtest or uptime monitor.

3. DataHUB of combined GitLab and InvenioRDM

The DataPLANT DataHUB as a science gateway is primarily based on the open-source framework GitLab. It provides an entry point to various services, starting with a versioned, generated web page and additional modules for community interaction. Key features of the DataHUB are versioning, the ability to work in groups, support for multiple contributions, and easy-to-use access management. The DataHUB platform is where the DataPLANT Annotated Research Contexts (ARCs) [4] evolve to a certain state. While these can be tagged or shared, GitLab is not intended to provide long-term access or citation. In order to fulfil these essential FAIR criteria, we have implemented a data publication service that complements the DataHUB. It is implemented using the turnkey repository framework InvenioRDM, supported by a large international community of research institutions and led by CERN in Switzerland. InvenioRDM was chosen over other frameworks due to the modern microservice-oriented design, the open-source nature of the project and the large community involved in the project. For our central DataHUB installation, we provide a separate Docker image with slightly different modifications. This image also includes the InvenioRDM publishing mechanism, which allows the user to publish an ARC directly to InvenioRDM from the DataHUB and automatically receive a DOI for the publication. To achieve this, we use GitLab's event hooking mechanism and a modified GitLab Auto DevOps pipeline. We prefer this approach to project-specific CI/CD templates so that users do not have to set them up themselves.

The goal of sustainability of the research data dictated our decision of underlying frameworks and technologies used in the DataPLANT infrastructure. The choice of a git-based solution for managing ARCs was driven by the widespread use of the protocol as well as predictable long-term adoption. While InvenioRDM has its own data model for publication, the community is already working on exporting datasets in platform-independent and well-established data packaging concepts such as OCFL and RO-Crates to enable potential future migrations.

Acknowledgement

We acknowledge the support of DataPLANT, funded by the German Research Foundation (DFG) within the framework of the NFDI – project number: 442077441. We thank the Ministry of Science, Research and Education in Baden-Württemberg for their support of the BioDATEN Science Data Center which provided the necessary means for the InvenioRDM workflow integration.

References

1. D. von Suchodoletz, T. Mühlhaus, J. Krüger, B. Usadel, and C. Martins Rodrigues "DataPLANT – ein NFDI-Konsortium der Pflanzengrundlagenforschung," 2021. [Online]. Available: <https://bausteine-fdm.de/article/view/8335>
2. T. Mühlhaus, D. Brillhaus, M. Tschöpe, O. Maus, B. Grüning, C. Garth, C. Martins Rodrigues, and D. von Suchodoletz, "DataPLANT – tools and services to structure the data jungle for fundamental plant researchers," in E-Science-Tage 2021: Share Your Research Data, V. Heuveline and N. Bisheh, Eds. Heidelberg: heiBOOKS, 2022, pp. 132–145. [Online]. Available: <https://doi.org/10.11588/heibooks.979.c13724>
3. D. von Suchodoletz, T. Mühlhaus, D. Brillhaus, H. Jabeen, B. Usadel, J. Krüger, H. Gauza and C. Martins Rodrigues, "Data Stewards as ambassadors between the NFDI and the community" in E-Science-Tage 2021: Share Your Research Data, V. Heuveline and N. Bisheh, Eds. Heidelberg: heiBOOKS, 2022, pp. 366–373. [Online]. Available: <https://doi.org/10.11588/heibooks.979.c13750>
4. C. Garth, J. Lukasczyk, T. Mühlhaus, B. Venn, J. Krüger, K. Glogowski, C. Martins Rodrigues, and D. von Suchodoletz, "Immutable yet evolving: Arcs for permanent sharing in the research data-time continuum," in E-Science-Tage 2021: Share Your Research Data, V. Heuveline and N. Bisheh, Eds. Heidelberg: heiBOOKS, 2022, pp. 366–373. [Online]. Available: <https://doi.org/10.11588/heibooks.979.c13751>

A FAIR Future for Engineering Sciences

Linking an RDM Community through a scientific journal

Izadora Silva Pimenta¹[\[https://orcid.org/0000-0001-7093-224X\]](https://orcid.org/0000-0001-7093-224X), Kevin T. Logan¹[\[https://orcid.org/0000-0001-5512-2679\]](https://orcid.org/0000-0001-5512-2679),
Michaela Leštáková¹[\[https://orcid.org/0000-0002-5998-6754\]](https://orcid.org/0000-0002-5998-6754) and Peter F. Pelz¹[\[https://orcid.org/0000-0002-0195-627X\]](https://orcid.org/0000-0002-0195-627X)

¹ Chair of Fluid Systems, Technische Universität Darmstadt, Germany

Abstract. The emergence of FAIR data management (FDM) is being witnessed in more and more disciplines, including the engineering sciences. However, until recently, little academic credit has been given for the work that sound FDM practices in research publications require. Moreover, there has been a lack of space where the engineering sciences community could discuss and share experiences, ideas and advice about this topic. In academia, a suitable platform for such information exchange are journals. In this publication, the concept behind *ing.grid*, the newly established open access journal for FDM in engineering sciences, is presented, illuminating how these challenges can be addressed by providing a platform for the publication of manuscripts, research data, and software as well as by incorporating open peer review.

Keywords: FAIR Data Management, Engineering Sciences, Open Science, Open Peer Review

1. Introduction

Producing and sharing research data, developing practical tools for processing that data and curating it to ensure it is findable, accessible, interoperable, and reusable (FAIR) as part of sound scientific practice is becoming more and more common in the engineering sciences. Scientists, funding associations as well as publishers increasingly recognise their importance [1]. FAIR data management (FDM) approaches, however, require considerable amount of time and experience. Hence, it is necessary to foster exchange between engineering scientists and help establish an FDM community among them, as well as giving scientific credit for the efforts made to ensure FDM.

2. Building a Community around FAIR Data Management

Encouraging a dialogue about FDM in engineering sciences can bring various opportunities, such as creating a mutual comprehension of core concepts, incorporating tried-and-tested solutions from fellow scientists and learning new workflows or best practices.

The journal *ing.grid*, the first scientific journal dedicated to FDM in engineering sciences, fosters collaborating and addressing the issues related to data management in this discipline and beyond whilst providing scientific credit for FDM. Established by active engineering scientists and librarians itself, *ing.grid* creates an environment where the community is connected and encourages to share and discuss research findings in an open peer review prior to as well as after publication.

3. With Openness towards a Stronger Community

ing.grid operates in accordance with fundamental strategies of the Open Science movement: Open Access and Open Peer Review [2]. Especially while establishing a new subject of scholarly research, vibrant scientific discussion and an engaged community are vital. Through its open peer review, ing.grid offers a platform for this discussion, making it transparent [3], accessible and inviting for the entire community. This also motivates reviewers to submit professional and high quality comments [3].

ing.grid developed a unique open peer review process, shown in Figure 1, that is based on the following key points:

- The manuscript publishing process is transparent. For each publication, readers can view the submitted version of the manuscript as well as the entire review discussion on ing.grid's own preprint server. Peer review no longer takes place behind closed doors.
- The high quality of publications is ensured by single-anonymised peer review. Once a manuscript is submitted, editors invite experts in the respective field to contribute review comments on the preprint repository, which are flagged as such.
- Both the scientific and non-scientific community can also participate in the open peer review process by sharing their comments on the platform. This fosters scientific exchange and community building. Editors supervising the publication process of a submission are obliged to moderate the community and review comments to ensure that the ing.grid community is a harassment-free experience for everyone.

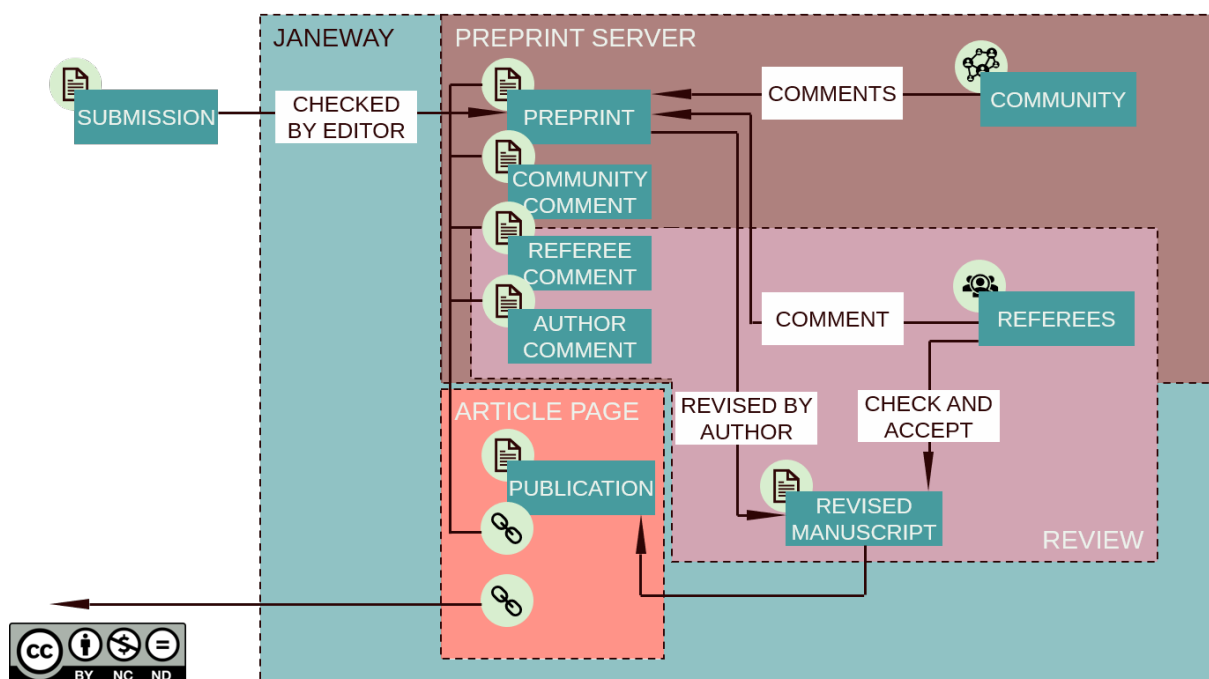


Figure 1. Open peer review process as implemented in ing.grid

Further promoting openness, ing.grid operates as a Diamond Open Access journal. Authors are not required to pay article processing charges, making it research-oriented rather than profit-oriented. All content is freely available to users and their institutions. The submissions are published under a CC-BY 4.0 license. This ensures that there are no barriers to joining the community.

4. Achieving FAIRness with a Journal

Most research activities in engineering sciences employ research data and research software. There is a trend towards making these FAIR [4, 5] and referencing them in the scientific publication. Some journals exist specifically for publishing research data [6] or research software [7]. To foster FAIRness, journals should provide infrastructure for connecting these three types of scientific output and for comprehensively linking them to one another.

ing.grid conceptualises all submissions as consisting of three components: manuscript, software and dataset. Depending on the submission type, one of the three is the primary component and the other two are supplementary material, Figure 2. ing.grid accepts three submission types: besides conventional research articles, it offers scientists the opportunity of having their software or dataset peer reviewed. While ing.grid only processes written text, it provides infrastructure to link the software and dataset repositories that are the subject of or supplementary material to its publications. This ensures the visibility of software and datasets, promoting their reuse.

Besides fostering FAIRness of research publications, ing.grid also considers itself to be a FAIR journal. It is Findable via its URL and will also be indexed in common journal databases in the future. As a Diamond Open Access journal, it is accessible. Interoperability is ensured through providing DOIs for the submissions of all types, and through accepting links to supplementary materials in other repositories. The concept behind ing.grid is also fully reusable as ing.grid is running on the open source platform Janeway.




	 manuscript	 software	 data
submission type			
mandatory material	manuscript	software descriptor link to software	data descriptor link to data
optional material	link to software link to data	link to data	link to software

Figure 2. Three submission types accepted by ing.grid: manuscripts, software descriptors and data descriptors along with mandatory and optional material (links to data and/or software).

5. Conclusion

To foster community engagement around FDM in engineering sciences, the journal ing.grid was founded. Strongly based on principles of Open Science and employing open peer review, ing.grid helps achieve FAIRness of scientific publications while being a FAIR journal itself.

ing.grid is a service developed by NFDI4ing that can be used within the whole NFDI and beyond. New initiatives such as the Data Literacy Alliance (DALIA) [8] can use the service as a platform for publishing scientific results while the concept of a scholar-led open access journal for FDM can be adopted also by other scientific disciplines.

As a journal, ing.grid bears responsibility for building an open science environment and increasing FAIRness [4] in Engineering Sciences. This way, ing.grid can ensure a FAIR future for scholarly communications [9].

Data availability statement

This submission is not based on data.

Author contributions

Izadora Silva Pimenta – Conceptualization, Visualization, Writing (original draft, review & editing)

Kevin T. Logan – Conceptualization, Visualization, Project Administration, Writing (original draft, review & editing)

Michaela Leštáková – Conceptualization, Visualization, Project Administration, Writing (original draft, review & editing)

Peter F. Pelz – Conceptualization, Funding acquisition, Project Administration, Supervision

Competing interests

The authors declare that they have no competing interests.

Funding

The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.

Acknowledgements

The authors would further like to thank the University and State Library Darmstadt, especially Thomas Stäcker, Anne-Christine Günther, Matthias Kerekes, Sebastian Kraußold and Gerald Jagusch, and the team behind the scholarly publishing platform Janeway, as well as Nils Preuß.

References

- [1] R. H. Schmitt *et al.*, "NFDI4Ing - the National Research Data Infrastructure for Engineering Sciences," 2020, doi: 10.5281/zenodo.4015201.
- [2] A. Capaccioni, "Open peer review: some considerations on the selection and management of reviewers," *JLIS.it*, vol. 14, no. 1, pp. 71–80, 2023, doi: 10.36253/jlis.it-508.
- [3] C. Wilcox, "Rude paper reviews are pervasive and sometimes harmful, study finds," *Science*, 2019, doi: 10.1126/science.aba5502.

- [4] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data*, early access. doi: 10.1038/sdata.2016.18.
- [5] M. Barker *et al.*, "Introducing the FAIR Principles for research software," *Scientific data*, early access. doi: 10.1038/s41597-022-01710-x.
- [6] Springer Nature Limited. "Journal Information | Scientific Data." <https://www.nature.com/sdata/journal-information> (accessed Apr. 19, 2023).
- [7] A. M. Smith *et al.*, "Journal of Open Source Software (JOSS): design and first-year review," *PeerJ preprints*, early access. doi: 10.7717/peerj-cs.147.
- [8] Fluidsystemtechnik – TU Darmstadt. "DALIA: Knowledge-Base für „FAIR data usage and supply“ als Knowledge-Graph." https://www.fst.tu-darmstadt.de/forschung_fst/zusammenarbeit_in_der_forschung/dalia/dalia_ueberblick.de.jsp (accessed Apr. 19, 2023).
- [9] Victoria Kitchener. "3 Ways scholarly journals can promote FAIR data." <https://blog.scholasticahq.com/post/ways-journals-can-promote-fair-data/> (accessed Apr. 19, 2023).

Conda, Container and Bots

How to build and maintain tool dependencies in workflows and training materials

Paul Zierep¹[\[https://orcid.org/0000-0003-2982-388X\]](https://orcid.org/0000-0003-2982-388X), Sanjay Kumar Srikakulam¹[\[https://orcid.org/0000-0002-1752-5060\]](https://orcid.org/0000-0002-1752-5060), Sebastian Schaaf^{1,2}[\[https://orcid.org/0000-0001-8193-0151\]](https://orcid.org/0000-0001-8193-0151), and Björn Grüning¹[\[https://orcid.org/0000-0002-3079-6586\]](https://orcid.org/0000-0002-3079-6586)

¹ Freiburg Galaxy Team, Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany

² ELIXIR Officers Team, Institute for Bio- and Geosciences 5, Research Center Jülich, Germany

Abstract. The lifecycle of scientific tools comprises the creation of code releases, packages and containers which can be deployed into cloud platforms, such as the Galaxy Project, where they are run and integrated into workflows. The tools and workflows are further used to create training material that benefits a broad community. The need to organize and streamline this tool development lifecycle has led to a sophisticated development and deployment architecture.

Additionally, it is crucial to keep the tools and their dependencies in sync, as well as to update downstream workflows and training material upon new tool releases. Outdated workflows and training material are disadvantageous to the community, but the effort to keep the tools up-to-date is an immense burden considering the huge amount of tools available. For example, there are more than 3200 tools hosted on the European Galaxy server (<https://usegalaxy.eu>) [1] as of this writing.

This talk will explain how the Galaxy community is automatizing the updates for tools, packages, containers, workflows, and training materials to lower the maintenance burden of the community.

The Galaxy Project [2] has made thousands of open-source software utilities for scientific data analysis easily accessible to researchers by providing computational infrastructure and a user-friendly web interface. Although Galaxy itself does not require any significant computational skills to use, the development and maintenance of new tools and workflows benefit from sophisticated infrastructure with both human and automated components. The process of integrating software into Galaxy requires knowledge of both the command-line interface of the underlying tool and the schema used by Galaxy to define interfaces in order to be able to write a “Galaxy tool wrapper”. Such a wrapper maps dataset inputs, additional parameters, and expected outputs.

The deployment of software tools envisioned for use in Galaxy is orchestrated by Conda [3]. Conda is a package-, dependency- and environment-manager that enables the distribution of software packages via dedicated channels (e.g. Bioconda, conda-forge). Currently, conda-forge and Bioconda provide more than 21,000 and more than 9,900 tools, respectively.

It is a community-driven project that aims to make it easier to install and manage a wide range of software on different platforms. Conda has several advantages over traditional methods of installing scientific software. It simplifies software installation by automatically managing dependencies and by providing a consistent environment for software development and analysis. The ease of use of Conda packages benefits the scientific community far beyond the Galaxy use cases. The success of this deployment strategy has already led to adaptation outside the bioinformatic field, e.g. in material science [4].

Bioconda [5] extends the basic Conda package management functionality by creating additional Docker and Singularity containers that provide a containerized tool environment [6]. Containerization is further promoted by the automatic generation of so called mulled containers, that provide different Conda dependencies in one container, for the use of tool wrappers that depend on multiple packages. Currently, there are more than 87,000 Docker [7] and Singularity [8] containers available that can be used in cloud infrastructures, HPC-, as well as on local compute environments.

The semi-automated development of the Galaxy tool wrappers is facilitated by the Planemo toolkit [9], which offers a wide range of functionalities for developing, deploying, and executing Galaxy tools, workflows, and training material, with a simple and powerful command-line interface. It also facilitates automated deployment of tools and automatic updates of software dependencies, and offers testing and linting functionality that helps to ensure high-quality tool wrappers and workflows. Although these tasks can be performed individually without Planemo, it provides a convenient single tool that encourages best practices and is already extensively used in the Galaxy ecosystem.

To ensure that tools, wrappers, workflows and the Galaxy servers are up-to-date, a system of bots is maintained that handle automated updates of components to their latest dependencies versions. The bots are based on Planemo as well as Bioconda utilities that are controlled by continuous integration (CI) GitHub pipelines. The demonstrated deployment architecture combines automated update functionality with necessary code-review steps in order to ensure quality controlled tools and wrappers. Updates of Bioconda packages, for example, will trigger pull requests (PRs) in dependent Galaxy tool repositories, but these PRs will require review and approval by expert developers before they can be merged.

Ideally, a bioconda tool update would propagate all the way to the main Galaxy servers, and to community-maintained workflows and training material using these tools.

In this talk, the ecosystem will be explained on the basis of a use case that follows one of the latest tools that were integrated into Galaxy. It will cover the complete tool development, testing and deployment process until its installation on the European Galaxy server and its integration into the Galaxy Training Network (GTN) [10] training material. Furthermore, the update propagation workflows will be demonstrated for this tool.

Keywords: conda, containers, bots, automation, Galaxy

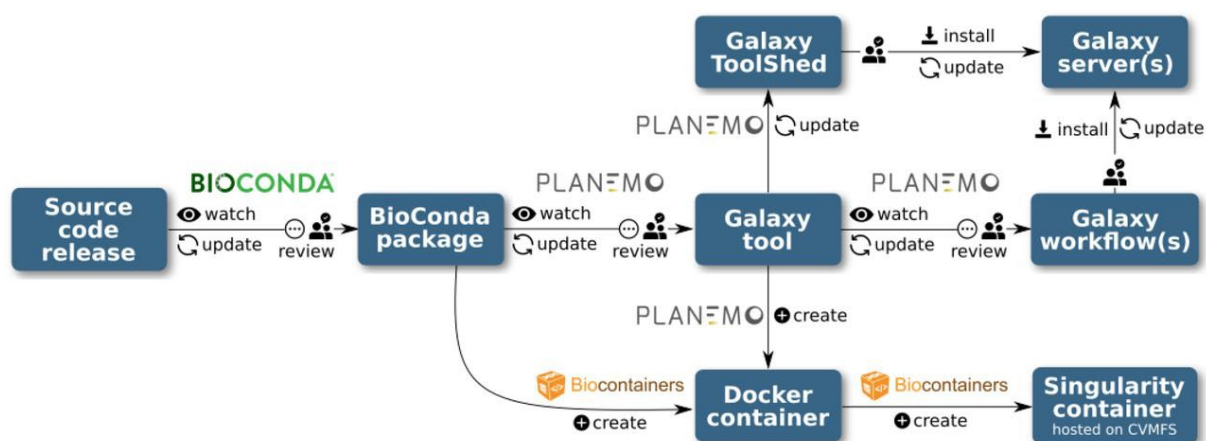


Figure 1. Overview flow diagram.

Competing interests

The authors declare that they have no competing interests.

References

1. "Galaxy Europe." <https://usegalaxy.eu> (accessed Apr. 24, 2023).
2. "Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update | Nucleic Acids Research | Oxford Academic." doi: 10.1093/nar/gkac247
3. "Conda — conda documentation." <https://docs.conda.io/en/latest/> (accessed Apr. 26, 2023).
4. J. Janssen et al., "pyiron: An integrated development environment for computational materials science," *Computational Materials Science*, vol. 163, pp. 24–36, Jun. 2019, doi: 10.1016/j.commatsci.2018.07.043.
5. B. Grüning et al., "Bioconda: sustainable and comprehensive software distribution for the life sciences," *Nat Methods*, vol. 15, no. 7, Art. no. 7, Jul. 2018, doi: 10.1038/s41592-018-0046-7.
6. F. da Veiga Leprevost et al., "BioContainers: an open-source and community-driven framework for software standardization," *Bioinformatics*, vol. 33, no. 16, pp. 2580–2582, Aug. 2017, doi: 10.1093/bioinformatics/btx192.
7. "Quay Container Registry · Quay." <https://quay.io/organization/biocontainers> (accessed Apr. 26, 2023).
8. "Index of /singularity/." <https://depot.galaxyproject.org/singularity/> (accessed Apr. 26, 2023).
9. S. Bray et al., "The Planemo toolkit for developing, deploying, and executing scientific data analyses in Galaxy and beyond," *Genome Res*, vol. 33, no. 2, pp. 261–268, Feb. 2023, doi: 10.1101/gr.276963.122.
10. "Galaxy Training," Galaxy Training Network. <https://training.galaxyproject.org/training-material/> (accessed Apr. 26, 2023).

Interactive Tools (IT) in Galaxy

Combining synchronous and asynchronous workflows

Tunc Kayikcioglu¹[\[https://orcid.org/0000-0003-2205-8062\]](https://orcid.org/0000-0003-2205-8062), Sanjay Kumar Srikakulam¹[\[https://orcid.org/0000-0002-1752-5060\]](https://orcid.org/0000-0002-1752-5060), Wolfgang Maier¹[\[https://orcid.org/0000-0002-9464-6640\]](https://orcid.org/0000-0002-9464-6640), and Björn Grüning¹[\[https://orcid.org/0000-0002-3079-6586\]](https://orcid.org/0000-0002-3079-6586)

¹ Freiburg Galaxy Team, Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany

Abstract. Galaxy [1] is a GUI-based scientific data management and analysis platform that aims to expand the target audience of scientific computing to scientists without advanced computer literacy or access to computing infrastructure. During the operation of a typical asynchronous workflow, the interaction of the user triggers a job submission to a hosted computing system, the results of which are then reported back to the user upon completion of the task. However, such an asynchronous interaction scheme is not compatible with scientific software that require user interaction during the execution itself (synchronous), e.g. many data visualisation software. Interactive Tools on Galaxy (IT; [2]) leverage tool and job components of Galaxy by taking advantage of the latest containerization solutions to provide access to interactive executable software. With ITs, Galaxy is enabling the more and more common use-case of asynchronous workflows while keeping the possibility to bridge at any time to synchronous workflows and reusing all the other Galaxy data management features.

Dynamic visual analytics applications, such as those using R Shiny, Dash or Ruby on Rails can be incorporated into Galaxy through an extension of the standard Galaxy tool framework. A developer needs to provide a container image (e.g., Docker) of the underlying application and a Galaxy tool specification (a.k.a. 'wrapper'). The existing job management system of Galaxy handles scheduling the execution of the container using its existing job management system. Interactive access through a built-in proxy is provided by presenting the user an auto-generated URL via the usual Galaxy GUI, which is forwarded to an internal port bound by the application of interest. The application will start in a user-specified configuration with access to input datasets. Users can then use the application interactively and commit the results of their work in the form of regular Galaxy output datasets. ITs can be seamlessly integrated into Galaxy: If the IT has been executed as part of a workflow, steps downstream of the IT will be scheduled once the expected outputs from the IT (whether generated through automated execution or through a human in the loop) become available. Irrespective of whether it is part of a Galaxy workflow, the saved efforts of the user will be recorded within their analysis history, and the exported 'notebooks' (the Jupyter 'scripts') can be shared with others as needed just as any other dataset. Such ITs provide a reproducible and resumable computing system which can leverage not only the computation infrastructure, but also the storage system.

As our scientific knowledge deepens, cross-disciplinary collaboration is becoming increasingly important for discovery, because it is rarely possible using only one set of skills. Galaxy currently supports sharing of research assets, such as workflows and histories, with other users and even with non-Galaxy users. However, these sharing capabilities are more limited to sharing a snapshot of the work rather than enabling virtual real-time collaboration. In contrast, the IT framework enables real-time visual analytics collaboration among authorized

users. As such, we expect ITs to have a similar enabling impact on collaborative teams as the multi-access office suite tools that make collaborative editing of research outputs possible. It is important to note that these simultaneous collaborative capabilities are being provided solely by our framework and do not require any modifications to the individual software for support at all.

As a case study, consider a user developing a simple machine learning model for the first time, a task involving a series of commands executed on some test data sets. Galaxy enables those exploratory research scenarios for example with the integration of Jupyter Notebooks [3][4], a popular open source web application that can be used to create and share documents containing source code and its products. A local deployment or usage of such an online resource would potentially present technical and security concerns and prohibit common usage scenarios such as institutional policies or hardware with restrictive configurations. Transfer of training data would also require the user to transfer large amounts of data over the network. The user would also likely need to purchase and/or configure a suitable graphics processor to be able to fully test her code and the model. ITs alleviate all of these issues as the users would immediately gain access to a pre-configured system with access to their other data in the Galaxy storage system.

Up to this date, our public instance, <https://live.usegalaxy.eu>, already had 20,000 IT jobs executed. We think that ITs can provide an opportunity to integrate much more interactive scientific use-cases and visualizations and will increasingly gain popularity in the community. In this talk we will walk through the integration and practical usage of the ITs with the aim of gradually increasing interest in deployment of new ITs and supporting more interesting and diverse use-cases.

Keywords: Interactive Tools, Galaxy, workflows, notebooks, Jupyter, R Studio

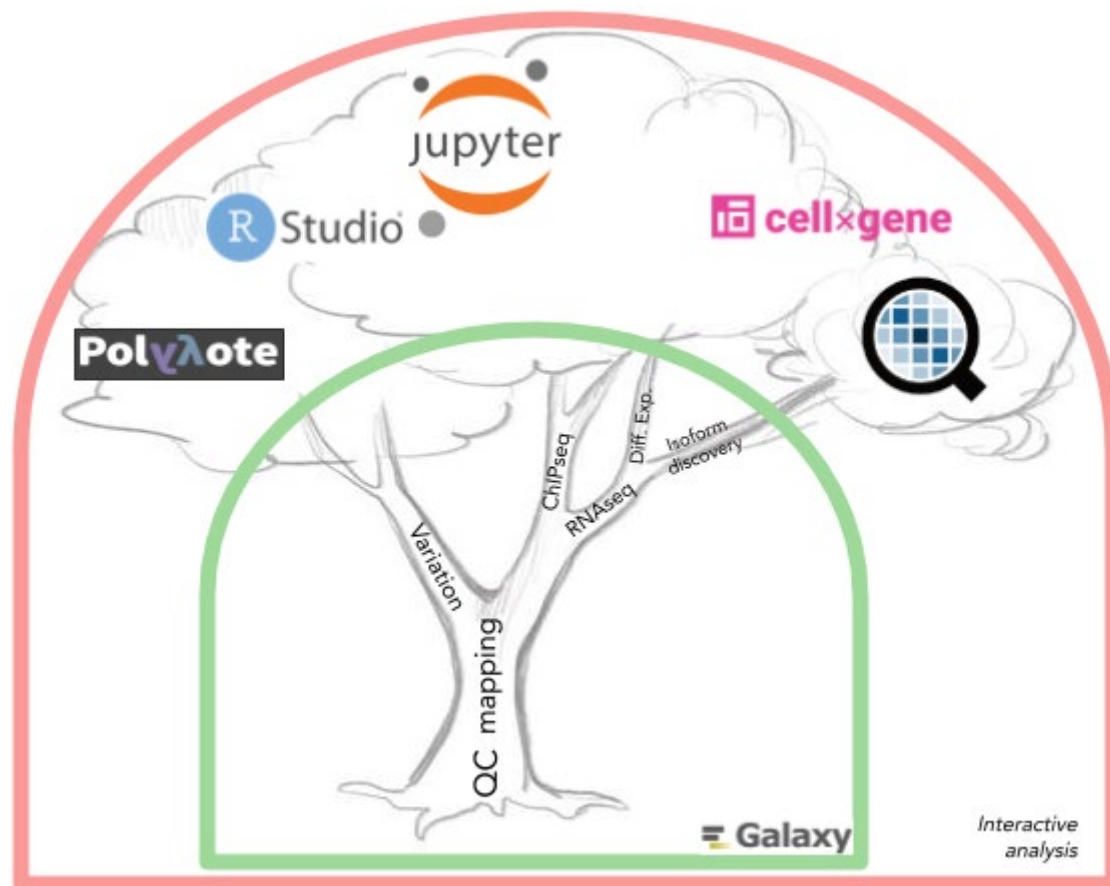


Figure 1. Galaxy and Interactive Tools (IEs). Galaxy’s interface works well for asynchronous processing of thousands of samples with existing tools (green outline). For exploratory data analyses, interactive environments are used. The aim is to enable integration between Galaxy and any IE such as programming frameworks such as Jupyter, RStudio or interactive visualization tools such as HiGlass. The user will be able to move the data between Galaxy and IEs transparently..

Competing interests

The authors declare that they have no competing interests.

References

1. The Galaxy Community - Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update; Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351, <https://doi.org/10.1093/nar/gkac247>
2. Gruning, B.A. et al. (2017) Jupyter and Galaxy: Easing entry barriers into complex data analyses for biomedical researchers. PLoS Comput Biol, 13, e1005425. <https://doi.org/10.1371/journal.pcbi.1005425>
3. <https://training.galaxyproject.org/training-material/topics/admin/tutorials/interactive-tools/tutorial.html> (accessed Apr. 24, 2023).
4. <https://realpython.com/jupyter-notebook-introduction/> (accessed Apr. 24, 2023).
5. <https://cfp.jupytercon.com/2020/schedule/presentation/184/climate-jupyterlab-as-an-interactive-tool-in-galaxy/> (accessed Apr. 24, 2023).

Quality Assessment for Research Data Management in Research Projects

Max Leo Wawer¹[\[https://orcid.org/0000-0003-3806-271X\]](https://orcid.org/0000-0003-3806-271X), Johanna Wurst¹[\[https://orcid.org/0000-0003-0430-5218\]](https://orcid.org/0000-0003-0430-5218),
Roland Lachmayer¹[\[https://orcid.org/0000-0002-3181-6323\]](https://orcid.org/0000-0002-3181-6323)

¹ Institute of Product Development, Leibniz University Hannover, Germany

Extended Abstract.

In the context of research data management (RDM), researchers are confronted with a multitude of new tasks and responsibilities. The totality of all tasks to ensure the re-use of data, long-term archiving, and access to data through data management planning, further data documentation, and provinces of data collection and analysis are described as research data management [1]. Often, the process of RDM is represented with data life cycle models, which include the basic phases of *planning*, *data collection*, *analysis*, *archiving*, *access*, and *reuse* [2].

When considering an engineering research methodology of the research process, it starts with the formulation of the research goals and planning of the research concept. This is followed by the analysis of empirical data based on data collection, processing, analysis, and interpretation to detail the as-is state. Based on the empirical analysis, the development of the solution for the improvement of the existing situation follows. This is followed by the evaluation of the solution through further analysis of empirical data concerning defined evaluation criteria [3]. In the field of engineering research, there are three main approaches to data collection: experiments, simulations, and theoretical analysis. These data are either collected by the researchers themselves or gathered from other sources and reused. This is followed by further processing and analysis of the data about properties of interest. A large number of such observations are carried out and published within a research project [4]. Research results that have been collected are published and stored continuously in a research project. Re-use of the data takes place in research during data collection in the sense of data collection. Integrating the activities and tasks of RDM into the research process results in the schematic process flow shown in Figure 1.

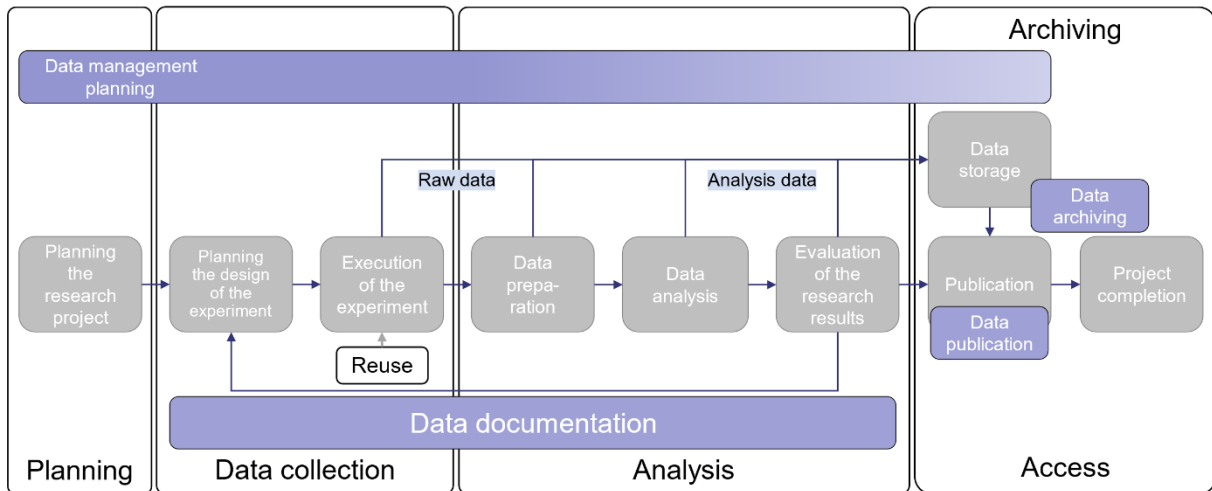


Figure 1. Integration of RDM into the engineering research process within a research project

In a research project, researchers are responsible for managing the data, adhering to standards of good scientific practice, preparing the data, and making it available and reusable for others throughout the life of the project. It is the researchers' responsibility to manage the data, develop contextual metadata and enable the re-use of the data. On the one hand, the quality of RDM in research projects must be ensured, and on the other hand, the quality of the data products must be guaranteed even after the end of the project. Fundamental to this is the traceability and guarantee of transparent research data [5]. In this context, guidelines for handling research data must be defined. The quality of the data always depends on the purpose and context of further processing, therefore the RDM should be oriented towards the research community and must be executed according to given standards to make comprehensible data available to the research community.

To ensure quality concerning the execution and implementation of processes, maturity models represent a method for qualitative evaluation. They enable an evaluation of objects and contents based on discrete maturity levels, from an initial to an optimized final state. Maturity models can be used to evaluate entire organizations or individual areas about defined strategic goals [6].

For the field of RDM, there are already developed maturity models that address RDM in various dimensions [7]. In these models, the RDM is considered as a whole system, and an evaluation of organizations, applications, and services is forced.

To assess the RDM in research projects, the NFDI4Ing [8] is developing maturity models oriented to the research process that enable researchers to assess the RDM independently during their research. The models focus on the research process-oriented execution of RDM towards a standardized and optimizing execution of RDM in research projects. The developed maturity characteristic of the maturity models (Figure 2) follow the contents of the Capability Maturity Model Integration (CMMI) and are aligned with the goals of RDM for execution in research projects. This ensures the traceability and integrity of research and RDM with increasing levels of maturity. The CMMI is an established maturity model that forms the basis for many developed maturity models.

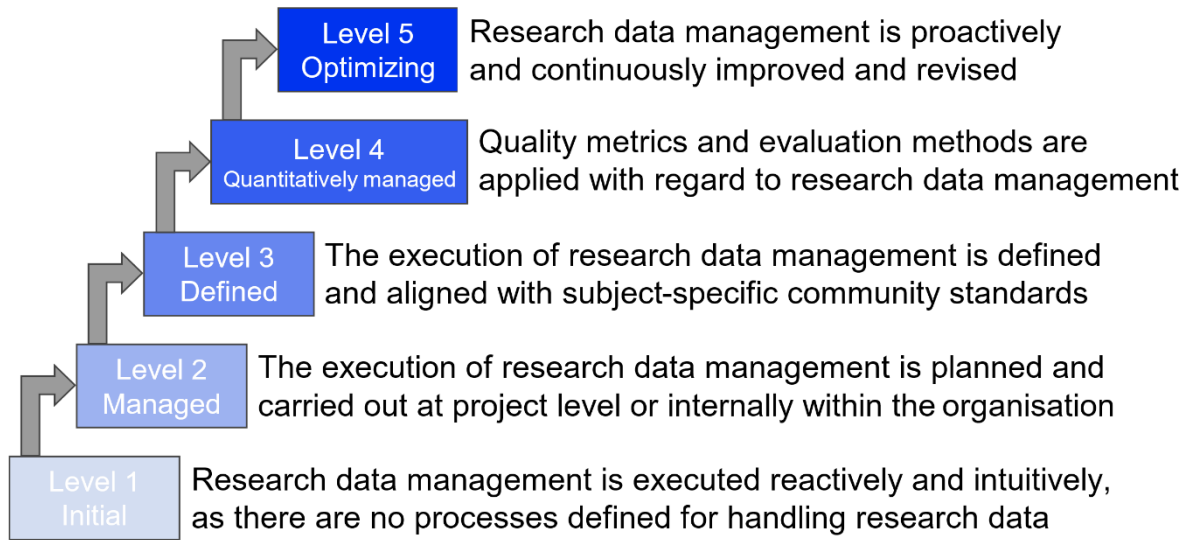


Figure 2. Developed maturity characteristic

The first level describes the execution of RDM in research projects, which does not follow any defined procedures. The RDM is not planned, but is done reactively and intuitively and depends on the commitment of the researchers. At the next level of maturity, RDM is planned and carried out in its defined areas at the project level. Within the project, the basic content for RDM is addressed and proactively executed. To publish comprehensible research data, the next stage is to align RDM with prevailing domain-specific community standards. This should ensure the interoperability and reusability of the research data in the respective research community. In level 4, content to ensure the quality of the research data and data management is then integrated into the processes. The last level provides a continuous and active improvement of the RDM solutions and active participation in the research community regarding the contents of the RDM.

To evaluate RDM in research projects and to improve it in perspective, individual maturity models are developed for the identified process areas of RDM (Figure 3) based on the developed maturity characteristic.

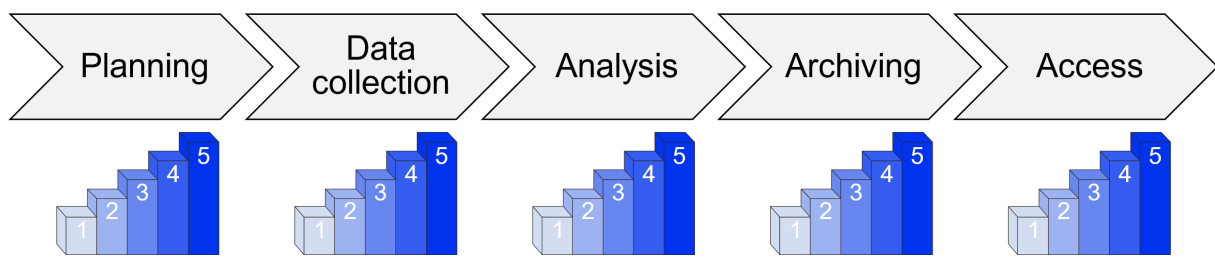


Figure 3. Defined process areas for the development of individual maturity models

In the individual maturity models, the contents and tasks of the process areas are taken into account and, based on this, goals and associated practices are defined at the individual maturity levels, oriented towards the maturity characteristics. The processes and activities of engineering research fields that have a strong impact on the area of data collection and analysis are taken into account [9]. In this way, the developed maturity models can be used to evaluate RDM in research projects uniformly and show possibilities for improvement toward standardized and secure RDM in the research community.

Keywords: maturity model, quality assessment, engineering research process, research data management

Acknowledgment

Max Leo Wawer, Johanna Wurst, and Roland Lachmayer would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713. "If you want to acknowledge persons or institutions you can do so here."

References

1. S. Buettner, H.-C. Hobohm, and L. Mueller. "Handbuch Forschungsdatenmanagement, 2011, 978-3-88347-283-6.
2. S. T. Kowalczyk, "Modelling the Research Data Lifecycle." IJDC 12, 2, 2017, pp. 331–361, doi: 10.2218/ijdc.v12i2.429.
3. L. T. M. Blessing and A. Chakrabarti. "Drm, a design research methodology," Springer, Heidelberg, 2014, 978-1-4471-5774-8.
4. D. Iglezakis and B. Schembera, "Anforderungen der Ingenieurwissenschaften an das Forschungsdatenmanagement der Universität Stuttgart - Ergebnisse der Bedarfsanalyse des Projektes DIPL-ING." o-bib 5, 3, 2018, pp. 46–60, doi: 10.5282/o-bib/2018H3S46-60.
5. M. Kindling, "Qualitätssicherung im Umgang mit digitalen Forschungsdaten / Quality assurance of digital research data / La garantie de la qualité des données numériques de recherche." Information - Wissenschaft & Praxis 64, 2-3, 2013, doi: 10.1515/iwp-2013-0020.
6. J. Becker, R. Knackstedt, and J. Pöppelbuß. "Developing maturity models for IT management: A procedure model and its application, 2009, doi: 10.1007/s12599-009-0044-5.
7. A. Lehmann and C. Odebrecht, "Reifegradmodelle im Forschungsdatenmanagement – IT-Prozessoptimierung im Wissenschaftsbetrieb." Information – Wissenschaft & Praxis 74, 1, 2023, pp. 9–21, doi: 10.1515/iwp-2022-2249.
8. R. H. Schmitt, V. Anthofer, S. Auer, S. Başkaya, C. Bischof, T. Bronger, F. Claus, F. Cordes, É. Demandt, T. Eifert, B. Flemisch, M. Fuchs, M. Fuhrmans, R. Gerike, E.-M. Gerstner, V. Hanke, I. Heine, L. Huebser, D. Iglezakis, G. Jagusch, A. Klinger, M. Krafczyk, A. Kraft, P. Kuckertz, U. Küsters, R. Lachmayer, C. Langenbach, I. Mozgova, M. S. Müller, B. Nestler, P. Pelz, M. Politze, N. Preuß, M.-D. Przybylski-Freund, N. Reißler-Pipka, M. Robinius, J. Schachtner, H. Schlenz, A. Schwarz, J. Schwibs, M. Selzer, I. Sens, T. Stäcker, C. Stemmer, W. Stille, D. Stolten, R. Stotzka, A. Streit, R. Strötgen, and W. M. Wang. "NFDI4Ing - the National Research Data Infrastructure for Engineering Sciences, 2020, doi: 10.5281/zenodo.4015200.
9. H. Dierend, O. Altun, I. Mozgova, and R. Lachmayer, "Management of Research Field Data Within the Concept of Digital Twin," In Advances in System-Integrated Intelligence, M. Valle, D. Lehnhus, C. Gianoglio, E. Ragusa, L. Seminara, S. Bosse, A. Ibrahim and K.-D. Thoben, Eds. Lecture Notes in Networks and Systems. Springer International Publishing, Cham, 2023, pp. 205–214, doi: 10.1007/978-3-031-16281-7_20.

Establishing Adaptive Governance in NFDI Consortia:

Lessons Learned from Deliberative Forums with Patients on their Role in the Governance of the German Human Genome-Phenome Archive (GHGA)

Eric Apondo¹, Andrea Züger¹, Andreas Bruns¹[\[https://orcid.org/0000-0002-1961-1691\]](https://orcid.org/0000-0002-1961-1691), Katja Mehlis¹, Christoph Schickhardt¹ and Eva Winkler¹

¹ National Center for Tumor Diseases, Heidelberg

Keywords: Governance, Consortia, Deliberative forums, translation, policy, personal health data, research

1. Introduction

There is widespread support for patient and public involvement (PPI) as an ethical requirement in biomedical research and policy development [1, 2], including research using personal health data, such as genomic data [3]. Sharing genomic and other personal health data is important for progress in health research [4], but poses ethical, legal and societal challenges for the governance of research and the institutions that conduct and support it. These challenges arise in part from the varied interests of the stakeholders involved, including patients, researchers, and funders [5]. In recent years, there has been a rise in collaborations between research institutions that conduct joint research programs. In these research consortia, the objectives and stakeholders in research are multiplied and more varied, making governance even more complex [6]. There has been debate about strengths and weaknesses of different forms of governance of research, with proponents of adaptive governance highlighting its responsiveness and flexibility to evolving goals and needs of stakeholders [5, 7]. However, there are no set standards for involving patients in the governance of health research. Moreover, as it is a highly context-specific process, cultural, legal and social contexts of the individual institutions or consortia must be considered.

Within the German National Research Data Infrastructure (NFDI), there are consortia supporting research with personal health data, one of which is the German Human Genome-Phenome Archive (GHGA) [8]. We describe GHGA's PaGODA Project (Patient Involvement in the Governance of an Omics Data Archive), whose goals are to gather patients' views for their involvement in the governance of GHGA by conducting deliberative forums, and to implement these views. We focus here on the process of translating the findings from the forums into the GHGA governance policy, and identify procedural factors that were important for the process.

2. The deliberative Forums

In this participatory project, we have collaborated with two patient communities (cancer and rare diseases), represented by patient experts as co-researchers in study design, writing the study protocol, developing discussion guides, recruitment, and writing results.

In July 2022, we conducted two one-day long, online, live deliberative forums with 26 members of the cancer and rare diseases communities in Germany (Table 1) on ethical issues related to the operations and governance of GHGA. Deliberative forums are a qualitative research method in which participants are educated about a complex issue, which they then discuss with a focus on dialogue and understanding varied points of view [9]. The quality of the deliberative forums was assessed using a participant questionnaire, as well as a pre- and post-survey instrument to measure knowledge gain and opinion shift.

3. Translation of the participants' recommendations into GHGA policy

The question arises: How can the preferences expressed by the patients be concretely implemented in GHGA's governance structure? The first step of the translation process was qualitative analysis of the forums, which was done using the framework approach [10, 11]. The results were summarized in a draft document in plain German, which was reviewed by GHGA members. The focus of this step was an assessment of the feasibility of participants' recommendations as well as their potential impact on other stakeholders, especially data controllers, researchers, and funders. The document was adjusted to include this assessment, then sent to the forum participants to make sure that the report reflected what was said in the forums. We invited the forum participants to a consensus-building dialogue event with GHGA members in March 2023 to discuss issues from the forums where there was a divergence of opinion. 17 participants attended. The resolutions from the dialogue event were included in the document, which, after a second round of feedback from GHGA members, became the final white paper on patient involvement in the governance of GHGA.

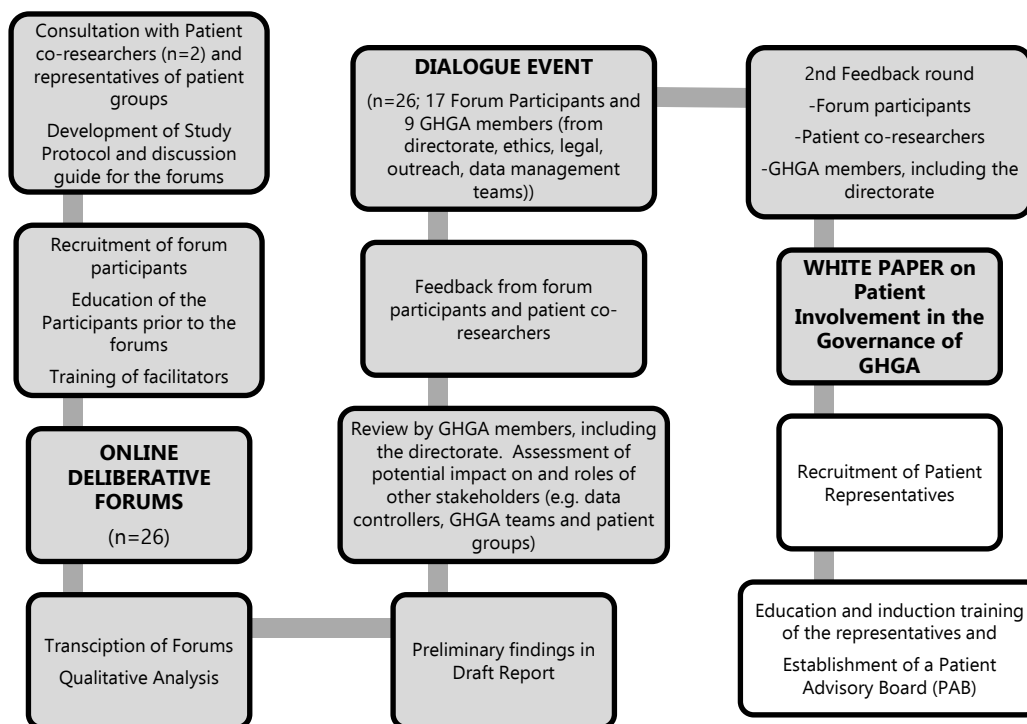


Figure 1. The steps of the PaGODA Participatory Project. The grey boxes indicate completed steps.

Table 1. Characteristics of participants of the dialogue forums.

Total number recruited	29
Did not participate due to illness	3
Forum participants	26
Male	8
Female	18
Disease category (more than one possible)	
Rare disease (RD)	10
Genetic predisposition for an RD	2
Cancer	7
Genetic predisposition for cancer	7
Relative of individual with RD	1
Representative of an RD Patient group	1
Age	
18-29	1
30-39	1
40-49	5
50-59	11
60-69	6
70-79	2
Level of Education	
Secondary school	6
Highschool diploma	6
College diploma	11
Doctorate	3

4. Results of the Translation Process

All topics on which there was a divergence of opinion among the participants were resolved during the dialogue event. Moreover, nearly all the overarching recommendations that were made by the forum participants were reflected or directly mentioned in the white paper. During the dialogue event, it was broadly agreed that a patient advisory board (PAB) consisting of patient representatives should be formed. A consensus was reached on the number of representatives, how they should be recruited, their roles, and whether they should be financially compensated. Concerning a recommendation pertaining to the GHGA data access procedures, potential impacts on data controllers were identified in the first feedback round. Possible roles of the data controllers in implementing the participants' recommendations were discussed and agreed upon in the dialogue event. Recommendations that were not reflected in the white paper were transparently discussed during the dialogue event to the satisfaction of all present.

5. Discussion and lessons learned: Key procedural factors that were necessary for implementation of participants' recommendations

Almost all the recommendations made by the patients in the deliberative forums will be reflected in the actual GHGA policy on governance. This is encouraging, as it has been previously observed that recommendations from deliberative exercises are rarely taken up into policy [12, 13]. Factors that contributed to this policy uptake [12] of the participants' recommendations into the white paper included: (i) Collaboration with patient co-researchers during the entire life-cycle of the project; (ii) Interdisciplinarity of the study team, which included expertise from ethics, medicine, philosophy, social sciences, and communications (iii) Attention to the framing of the goals of the project and of deliberation topics, taking into account legal and procedural constraints, and communicating these clearly to the participants; and (iv) Support of GHGA members and the leadership in feedback rounds and impact assessment. These factors should be considered by personal health data infrastructures and consortia aiming to implement adaptive governance frameworks that are responsive to the needs and perspectives of their stakeholders.

Competing interests

The authors have no conflicts of interest to declare.

Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of GHGA – The German Human Genome-Phenome Archive (www.ghga.de, Grant Number 441914366 (NFDI 1/1))

References

1. Kaye J, Terry SF, Juengst E, Coy S, Harris JR, Chalmers D, et al. Including all voices in international data-sharing governance. *Human Genomics*. 2018;12(1):13. doi: 10.1186/s40246-018-0143-9
2. McCoy MS, Warsh J, Rand L, Parker M, Sheehan M. Patient and public involvement: Two sides of the same coin or different coins altogether? *Bioethics*. 2019;33(6):708-15. doi: 10.1111/bioe.12584
3. Jasanoff S, Hurlbut JB, Saha K. Democratic Governance of Human Germline Genome Editing. *Crispr j*. 2019;2(5):266-71. doi: 10.1089/crispr.2019.0047

4. Global Alliance for Genomics and Health (GA4GH). Enabling responsible genomic data sharing for the benefit of human health 2023 [Available from: www.ga4gh.org. 10 Apr 2023
5. Juengst ET, Meslin EM. Sharing with Strangers: Governance Models for Borderless Genomic Research in a Territorial World. *Kennedy Institute of Ethics Journal*. 2019;29:67 - 95. doi:
6. Pratt B, Hyder AA. Governance of Transnational Global Health Research Consortia and Health Equity. *Am J Bioeth*. 2016;16(10):29-45. doi: 10.1080/15265161.2016.1214304
7. Brunner RD, Steelman TA. Towards Adaptive Governance. In: Brunner RD, Steelman TA, Coe-Juell L, Cromley CM, Edwards CM, Tucker DW, editors. *Adaptive Governance: Integrating science, policy and decision making*. New York: Columbia University Press; 2005.
8. National Research Data Infrastructure (NFDI) e.V. The German National Research Data Infrastructure (NFDI) Karlsruhe: NFDI; 2023 [Available from: <https://www.nfdi.de/association/?lang=en>. 10 Apr 2023
9. Goodin RE. *Innovating democracy: Democratic theory and practice after the deliberative turn*. Oxford, UK: Oxford University Press; 2008.
10. Pope C, Ziebland S, Mays N. Analysing Qualitative Data. *BMJ*. 2000;320:114-6. doi: 10.1136/bmj.320.7227.114
11. Nicol D, Critchley C, McWhirter R, Whitton T. Understanding public reactions to commercialization of biobanks and use of biobank resources. *Soc Sci Med*. 2016;162:79-87. doi: 10.1016/j.socscimed.2016.06.028
12. O'Doherty K, Hawkins A. Structuring Public Engagement for Effective Input in Policy Development on Human Tissue Biobanking. *Public health genomics*. 2010;13:197-206. doi: 10.1159/000279621
13. Street J, Duszynski K, Krawczyk S, Braunack-Mayer A. The use of citizens' juries in health policy decision-making: A systematic review. *Social Science & Medicine*. 2014;109:1-9. doi: 10.1016/j.socscimed.2014.03.005

Developing Consent Tools for the Research Community at the German Human Genome-Phenome Archive (GHGA)

Andreas Bruns¹[\[https://orcid.org/0000-0002-1961-1691\]](https://orcid.org/0000-0002-1961-1691), Simon Parker^{2,3}[\[https://orcid.org/0000-0001-9993-533X\]](https://orcid.org/0000-0001-9993-533X),
Fruzsina Molnár-Gábor²[\[https://orcid.org/0000-0002-9406-2776\]](https://orcid.org/0000-0002-9406-2776), Eva C. Winkler¹[\[https://orcid.org/0000-0001-7460-0154\]](https://orcid.org/0000-0001-7460-0154),
and the GHGA Consortium

¹ University Hospital Heidelberg, National Center for Tumor Diseases (NCT), Section Translational Medical Ethics

² Universität Heidelberg, BioQuant, Governance of Emerging Technologies in Medical Research and Health Care

³ German Human Genome-Phenome Archive (GHGA, W620), German Cancer Research Center (DKFZ)

Abstract. The German Human Genome-Phenome Archive (GHGA) aims to enable the responsible sharing of human omics data for secondary research use across Germany and Europe. Informed consent is the most commonly used legal and ethical basis for processing omics data for secondary use. However, obtaining informed consent from Data Subjects can be challenging when data is to be widely shared and reused beyond the initial purpose of collection. To address these challenges, the ELSI (Ethical, Legal, and Social Implications) Group of GHGA has developed consent tools for the research community. First, we have developed a toolkit for prospective data collection, which consists of consent modules and complementary advice on how to update or create new consent forms. Second, we have created a legacy consent toolkit that can be used by researchers to assess whether the consent under which data was originally collected covers further data processing for secondary research purposes.

Keywords: Informed consent, Legacy consent, GDPR, Omics data, Secondary use

1. Introduction

The German Human Genome-Phenome Archive (GHGA) is currently establishing a federated data infrastructure allowing the secure storage of, and controlled access to, human genome and other omics data for scientific research use. So long as omics data relates to a specific identifiable person (Data Subject), it is considered sensitive personal data. Informed consent is the most commonly used legal and ethical basis for processing such data for scientific research purposes. Using consent as a legal basis has several advantages; it is a legal basis for processing both personal and special category personal data, and its implementation enables data consented for research use to be shared across European nations easily. Outside of its legal value, informed consent is a central ethical principle as consent ensures that Data Subjects are not treated as mere means to the ends of research but as participants freely choosing to contribute to research activities.

Importantly, consent is needed not only for collecting and using omics data for the primary purpose but also for sharing data for secondary research use. However, obtaining consent from data subjects for secondary use can be challenging. According to the European Union's General Data Protection Regulation (GDPR (EU) 2016/679), consent should be freely

given, unambiguous, specific, and informed. While in biomedical research with derivatives such as medical data there has been a shift away from strictly project-specific consent to so-called broad consent (consent to a broadly defined range of future research uses), it can be difficult to ensure that consent is sufficiently informed when data is to be widely shared and reused beyond the initial purpose of collection. Moreover, researchers may wish to process data collected prior to the introduction of the GDPR and may wonder whether the original consent allows data sharing and secondary use.

GHGA seeks to help address these regulatory challenges. The ELSI (Ethical, Legal, and Social Implications) Group of GHGA has developed consent tools for the research community to help researchers ensure that data is collected and shared with the consent of Data Subjects. First, we have developed a Modular Consent Toolkit for prospective data collection, which consists of consent modules and complementary advice on how to update or create new consent forms. Second, we have created a Legacy Consent Toolkit for legacy data (previously collected data) to assess whether the consent under which data was originally collected covers further data processing for secondary research purposes.

2. The GHGA Consent Tools

2.1. The Modular Consent Toolkit

The Modular Consent Toolkit, which contains consent modules and complementary guidance on how to use them, can be used to update or create new consent forms to enable data sharing for secondary research use and is available as a white paper on Zenodo under open access [1]. The consent modules explain the relevant processes around data sharing and secondary use, using comprehensible yet legally appropriate language, and are specifically designed to enable broad future research use and archival of the data being collected.

There is a total of four consent modules, i.e., text blocks that can be integrated into consent form text. All modules are available in German and in English. (1) The central module, the Data Sharing Module, explains what happens when data is securely archived and made available to researchers for secondary research use under conditions of controlled access. (2) The De-Identification Module provides further information on the process of removing direct identifiers and the security status of de-identified data. (3) The Controlled Access Module elaborates on the process of how researchers can request and be granted access to data. (4) Finally, the Consent Options Module allows Data Subjects to record their consent decision regarding the secondary research use of their data. Following the consent modules and guidance on how to use them, the white paper also contains a section on how researchers are to evaluate consent forms that have been updated using these consent modules to ensure that these are consistent and coherent.

2.2. The Legacy Consent Toolkit

The Legacy Consent Toolkit addresses the issue of legacy data and can be used to assess whether the consent under which data was originally collected allows further data processing for secondary research purposes. It can be accessed via the GHGA Website [2]. The Toolkit does not provide formal legal advice on whether legacy consent is a legitimate legal basis for processing but is instead used to provide guidance.

The structure of the Legacy Consent Toolkit is designed to ascertain whether the proposed secondary processing increases the risk to the rights and freedoms of the Data Subjects in a manner unforeseen to the Data Subjects at the time they gave their consent. To do so, it is necessary to compare the primary and proposed secondary processing. In the first stage of the assessment, the information content at the time of the secondary processing is compared to that originally collected. In the second stage, the purpose of the primary and secondary

processing is compared. In the third stage, the person originally permitted to process the data is compared to the proposed secondary processor. In the final stage, an assessment is performed to understand whether the secondary processing increases the risk to the Data Subjects from a data protection perspective.

3. Conclusion

While our consent toolkits provide a service to researchers and their institutions hoping to ensure consent for the sharing of omics data for scientific research use, they may also function as a measure to protect the interests of patients and other groups of Data Subjects. So long as consent remains the basis for data sharing for secondary research use, it is important that Data Subjects are put in a position to make informed decisions about the further use of their data. Our toolkits aim to help Data Subjects gain a better understanding of the implications of secondary use as well as to ensure that it is properly assessed whether further use is covered by their original provision of consent.

Data availability statement

The submission is not based on any data.

Underlying and related material

None.

Author contributions

Andreas Bruns and Simon Parker have shared first authorship for this abstract ("Writing – original draft", "Writing – review and editing"). Fruzsina Molnár-Gábor and Eva C. Winkler have supervised the research on which this abstract reports ("Supervision"). The GHGA Consortium functions as collective co-author on all GHGA publications and has acquired funding for this research ("Funding acquisition").

Competing interests

The authors declare that they have no competing interests.

Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of GHGA – The German Human Genome-Phenome Archive (www.ghga.de, Grant Number 441914366 (NFDI 1/1)).

Acknowledgement

The authors would like to thank the GHGA Team for valuable feedback on the Consent Tools.

References

1. Andreas Bruns, Anna Benet-Pages, Jan Eufinger, Holm Graessner, Oliver Kohlbacher, Fruzsina Molnár-Gábor, Simon Parker, Christoph Schickhardt, Oliver Stegle, & Eva

Winkler, "Consent Modules for Data Sharing via the German Human Genome-Phenome Archive (GHGA)," Zenodo, July, 2022, doi: 10.5281/zenodo.6828131, <https://zenodo.org/record/6828131#.ZEKD3i0Rp-V> (21/04/2023).

2. GHGA, "Legacy Consent Toolkit." GHGA Website. <https://www.ghga.de/resources/legacy-consent> (21/04/2023).

Distributed Privacy-Preserving Data Analysis in NFDI4Health with the Personal Health Train

Yongli Mou¹[\[https://orcid.org/0000-0002-2064-0107\]](https://orcid.org/0000-0002-2064-0107), Feifei Li²[\[https://orcid.org/0000-0003-4815-8547\]](https://orcid.org/0000-0003-4815-8547), Sven Weber²[\[https://orcid.org/0000-0002-8518-9097\]](https://orcid.org/0000-0002-8518-9097), Sabith Haneef³, Hans Meine¹[\[https://orcid.org/0000-0002-7557-5007\]](https://orcid.org/0000-0002-7557-5007), Liliana Caldeira¹[\[https://orcid.org/0000-0002-9530-5899\]](https://orcid.org/0000-0002-9530-5899), Mehrshad Jaberansary²[\[https://orcid.org/0000-0003-3407-1387\]](https://orcid.org/0000-0003-3407-1387), Sascha Welten¹[\[https://orcid.org/0000-0001-5570-9672\]](https://orcid.org/0000-0001-5570-9672), Yeliz Yediel Ucer^{1,3}[\[https://orcid.org/0000-0002-6845-7774\]](https://orcid.org/0000-0002-6845-7774), Guido Prause¹[\[https://orcid.org/0009-0008-4273-4957\]](https://orcid.org/0009-0008-4273-4957), Stefan Decker^{1,3}[\[https://orcid.org/0000-0001-6324-7164\]](https://orcid.org/0000-0001-6324-7164), Oya Beyan^{2,3}[\[https://orcid.org/0000-0001-7611-3501\]](https://orcid.org/0000-0001-7611-3501), and Toralf Kirsten^{4,5}[\[https://orcid.org/0000-0001-7117-4268\]](https://orcid.org/0000-0001-7117-4268)

¹Chair of Computer Science 5, RWTH Aachen University, Germany

²Institute for Medical Informatics, Faculty of Medicine, University Hospital Cologne, University of Cologne, Germany

³Fraunhofer FIT, Sankt Augustin, Germany

⁴Department of Medical Data Science, University Medical Center Leipzig, Germany

⁵Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Germany
Fraunhofer MEVIS, Bremen, Germany

Department of Diagnostic and Interventional Radiology, Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany

Abstract: Data sharing is often met with resistance in medicine and healthcare, due to the sensitive nature and heterogeneous characteristics of health data. The lack of standardization and semantics further exacerbate the problems of data fragments and data silos, which makes data analytics challenging. NFDI4Health aims to develop a data infrastructure for personalized medicine and health research and to make data generated in clinical trials, epidemiological, and public health studies FAIR (Findable, Accessible, Interoperable, and Reusable). Since this research data infrastructure is distributed over various partners contributing their data, the Personal Health Train (PHT) complements this infrastructure by providing a required analytics infrastructure considering the distribution of data collections. Our research has demonstrated the capability of conducting data analysis on sensitive data in various formats distributed across multiple institutions and shown great potential to facilitate medical and health research.

Keywords: NFDI4Health, distributed data analytics, personal health train

1 Introduction

In the medical sciences, data is continuously collected from patient care services, registries, clinical trials, epidemiologic studies, and other research projects. Data collections are managed in a highly fragmented manner. Only a few of them are web-based accessible, while most of them are not findable and often there is no widespread knowledge about them - even within the same institution where the study was conducted. The NFDI4Health consortium, as part of the German National Research Data Infrastructure (NFDI) Initiative, aims at bridging the highly fragmented data collections generated by clinical trials and epidemiological and public health studies. To overcome the current limitations resulting from this fragmentation, NFDI4Health establishes an infrastructure allowing each data owner to continually manage its data collection locally and, thus, keep the sovereignty about the data, but need to make the data collection FAIR (Findable, Accessible, Interoperable, Reusable) and, hence, register it at a so-called Local Data Hub (LDH). There are LDHs all over Germany that are connected with a German Central Health Study Hub (CSH) which is the central access point for scientists to search and request data of interest. To complement the distributed data management, NFDI4Health contributes with infrastructures allowing to analyse data in a distributed mode, such as the Personal Health Train (PHT). In this paper, we sketch how the PHT is used within NFDI4Health to share medical study data for a common analysis and to simultaneously preserve the privacy of personalized medical data.

2 Methods and Materials

The predominantly used centralized analysis requires all requested data to be collected from partners who want to contribute to the medical research. On the contrary, the PHT constitutes a paradigm shift, i.e., bringing the algorithm to the data, and provides a novel distributed flexible approach that enables the use of sensitive personal data for privacy-preserving data analysis in a network of participants, while data owners stay in control of their own data [1]. Incorporating the FAIR principles, the PHT aims to facilitate medical and health research that applies not only to populations but also can be tailored to individuals.

The PHT has two main concepts, namely the Stations and Trains [2]. The stations are nodes that provide computational resources and execute analytic tasks in a secure environment in a way that it can run without any further installation. Specific analysis tasks are encapsulated and executed at stations. A station will be attached to an LDH, and data will be attached (connected to) at each station. One of the primary motivations behind the PHT is to empower data owners to take control of their health data. Therefore data remain in their original location attached at an LDH, and there is no automatic execution of the overall train keeping the opportunity for each data provider to check the source code before and the obtained results from their own station afterward.

Beyond these advantages, the most important feature of the PHT is the flexibility of the choice of data source technologies, as it can deal with different data types (e.g., radiology or genomics) and data formats or standards (e.g., FHIR or DICOM) [3], [4]. Besides, the PHT is agnostic to the code languages such that researchers can develop their own algorithms in different programming languages, including R and Python.

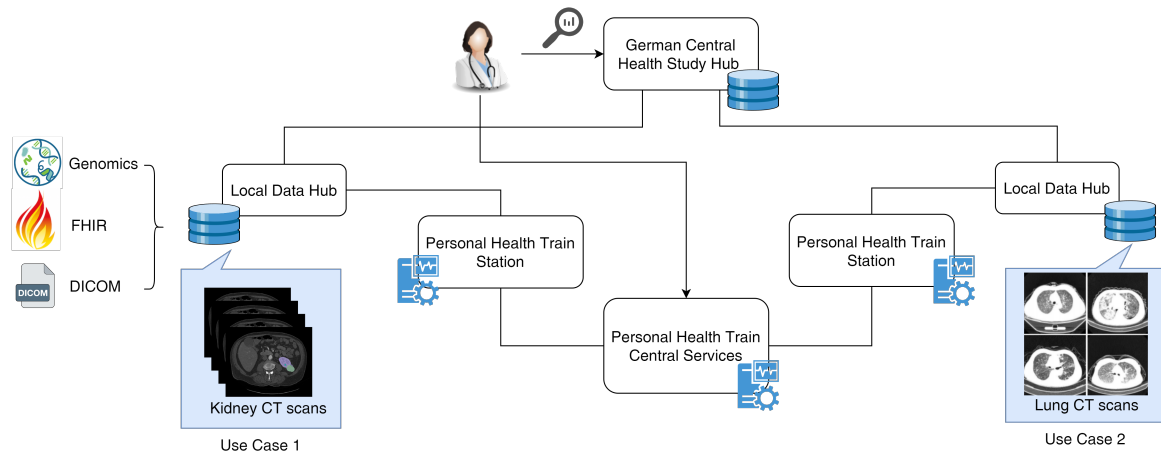


Figure 1. In NFDI4Health, the decentralized architecture comprises the German Central Health Study Hub (CSH) and a collection of Local Data Hubs (LDHs) and the Personal Health Train (PHT) provides the infrastructure for distributed analytics. LDHs are connected with the CSH and a PHT station is attached to an LDH. Researchers search and request data of interest through CSH and request data analytics through the PHT central services.

3 Results

While different groups in NFDI4Health work highly parallel on several infrastructural components and organizational procedures (e.g., governance acts) for the overall architecture, it is challenging to provide infrastructural components without precise requirements and interface definitions. Therefore, we are in contact with some groups within NFDI4Health while improving, testing, and adapting the PHT analysis infrastructure to new requirements and interface designs. While the first set of LDHs is currently established and filled with first content, i.e., metadata about clinical and epidemiological studies, there is currently a low but growing number of studies accessible in this early project stage that can be included in typical data analysis. However, at this project stage, data that can be used to solve medical research problems using complex algorithms, such as from the artificial intelligence method spectrum, or of more complex types, such as image and genomics data, is still missing. Therefore, we created use cases to show that the PHT infrastructure is working in principle, as shown in Figure 1.

A first use case focuses on the recognition of kidney tumors in patients to characterize the current stage of the disease and the location of the tumor within the kidney in order to provide therapy recommendations. The basis for this study is computer tomography (CT) image collection of known tumor patients in two locations. While these patients have already been treated in hospitals, the idea is to include them in a multi-center study to evaluate the overall segmentation, which can later be used for assessing the outcome of different therapy approaches. Recognizing the tumor stage and location requires segmenting the available CT images and identifying the tumor portion(s) within the images. We apply methods from the deep learning spectrum, in particular, the nnUNet [5] for 3D segmentation and then use this to extract the radiomics features. Instead of moving CT images for a centralized analysis, the PHT circulates the trained nnUNet model from server to clients and then transfers the extracted features for a federated analysis.

A second use case focuses on recognizing lung cancer in patients available at multiple sites. However, image data need to be harmonized when they have been taken by different devices and/or by different protocols. The amount of images produced by each device and protocol is different and sometimes small. Therefore, the approach is to produce synthetic data taking the available CT images into account in a way the synthetically generated data amount is equally distributed over devices and protocols and large enough for learning the differences from each subtype. The synthesized data is generated based on the model CycleGAN [6] at each LDH. It is used for style transfer and synthetic data generation to mitigate the distribution shifts from different devices and protocols. Based on this generated synthetic data set, we will apply a 3D pre-trained segmentation model [7], allowing us to recognize and extract the anatomical structures within the lung.

4 Discussion

In response to the growing demand for more extensive knowledge acquisition through improved utilization of research data and the ensuing social advantages, the PHT offers the necessary infrastructure to facilitate secure, privacy-preserving, and standardized distributed data analytics across various medical and health data providers and researchers. Our studies have made significant progress in integrating federated and incremental learning using the PHT infrastructure. Taking both use cases into account, we will study future challenges of distributed analysis, in particular, when complex analysis methods are applied such as from machine learning method spectrum or by using complex data types including large-scale images (e.g., magnetic resonance images, CT) and genetics data. Moreover, we will align the requirements obtained from the intended use cases with those from other NFDI4Health groups in terms of interoperability, interrelating PHT stations to LDHs and the CSH as well as governmental procedures.

Funding

This work was done as part of the NFDI4Health Consortium (www.nfdi4health.de). We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 442326535.

References

- [1] O. Beyan, A. Choudhury, J. Van Soest, *et al.*, “Distributed analytics on sensitive medical data: The personal health train,” *Data Intelligence*, vol. 2, no. 1-2, pp. 96–107, 2020.
- [2] S. Welten, Y. Mou, L. Neumann, *et al.*, “A privacy-preserving distributed analytics platform for health care data,” *Methods of information in medicine*, vol. 61, no. S 01, e1–e11, 2022.
- [3] Y. Mou, S. Welten, M. Jaberansary, *et al.*, “Distributed skin lesion analysis across decentralised data sources,” in *Public Health and Informatics*, IOS Press, 2021, pp. 352–356.
- [4] S. Welten, L. Hempel, M. Abedi, *et al.*, “Multi-institutional breast cancer detection using a secure on-boarding service for distributed analytics,” *Applied Sciences*, vol. 12, no. 9, p. 4336, 2022.
- [5] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation,” *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

- [6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [7] J. Wasserthal, M. Meyer, H.-C. Breit, J. Cyriac, S. Yang, and M. Segeroth, "Totalsegmentator: Robust segmentation of 104 anatomical structures in ct images," *arXiv preprint arXiv:2208.05868*, 2022.

“Hello ELSA, how are you?”

Legal and ethical challenges in RDM, current and future tasks of ELSA activities against the background of AI and Anonymisation

Franziska Boehm¹, Ulrich Sax² [<https://orcid.org/0000-0002-8188-3495>], Oliver Vettermann³ [<https://orcid.org/0000-0001-7393-1103>], Paweł Kamocki⁴ [<https://orcid.org/0000-0003-4881-7549>] and Vasilka Stoilova⁵ [<https://orcid.org/0000-0002-8961-509X>]

¹ FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Karlsruhe; Karlsruhe Institute for Technology; NFDI4Culture.

² Department of Medical Informatics, University Medical Center, Göttingen; NFDI4Health.

³ FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Karlsruhe; NFDI4Culture.

⁴ Leibniz Institute for the German Language, Mannheim; Text+.

⁵ University Library, Mannheim; BERD@NFDI.

Abstract. The proposed contribution will shed light on current and future challenges on legal and ethical questions in research data infrastructures. The authors of the proposal will present the work of NFDI’s section on Ethical, Legal and Social Aspects (hereinafter: ELSA), whose aim is to facilitate cross-disciplinary cooperation between the NFDI consortia in the relevant areas of management and re-use of research data.

Keywords: NFDI section; legal aspects; anonymisation; artificial intelligence; language models.

1. Introduction

Regarding the progress made in ELSA, we would like to present first results which are the outcomes of a workshop on anonymisation and of the ELSA Task Force aiming to put them forward for discussion. We discussed methods and tools around the anonymization of data spanning from tabular data [1] to free text data [2]. Another task force develops a decision tree on how to tackle privacy questions in research projects. Furthermore, we would like to engage in discussions on proposals for cross-consortium guidelines and legal standards. Specific topics we would like to discuss are data protection law with a focus on anonymisation as well as intellectual property law with a focus on legal status of AI-generated data. We also like to bring forward related ethical implications of using generative AI tools which were already discussed within ELSA.

2. Results

Moreover, new insights on questions arising out of the challenges of anonymisation and AI in a research data management context are demonstrated. Such challenges will affect all projects that (a) collect, process, or archive (personal) research data and/or (b) work with data or artefacts in which persons or organisations hold exploitation rights. By drawing on existing experi-

ence, we aim to enter into dialogue with rights holders in order to promote effective data protection and adhere to the research ethics guidelines (including the DFG principles of good scientific practice). Additional legal obstacles may arise from protecting research results that incur third-party rights. Solutions are to be prepared from various specialist communities. Lastly, ELSA also outlines training content for initial, further and continuing training in cooperation with section Training & Education (short: edutrain), as well as to involve external stakeholders in those areas where it seems appropriate. Addressing the privacy concerns and the accompanying legal uncertainties of NFDI consortia is the main goal of the newly created ELSA Task Force Data Protection. Task Force members focus on drafting standardized guidelines, which will enable researchers to adopt a FAIR approach to RDM while also protecting the personal data of their research subjects and adhering to the data protection regulations.

The rise of generative AI tools, such as Midjourney or ChatGPT, and their rapid development also created unexpected challenges for research data management, which were addressed by ELSA members. Especially in the field of text data, the so-called Large Language Models (or LLMs) like GPT-4 were a real gamechanger. Generative AI outputs are of increasing value for research, and are used as research data. Their legal status, in particular with regard to intellectual property rights, is difficult to determine, as illustrated by the recent Statement of Policy of the US Copyright Office (USCO). According to USCO, the mere fact of prompting an AI tool is not enough to claim copyright in the output. Still, works containing AI-generated material can be eligible for copyright protection, e.g. if such material is modified or arranged by a human author. In practice, the distinction is extremely difficult to make, especially that currently there is no reliable method to detect texts generated by ChatGPT.

Another pressing issue related to research data management and AI is to what extent data collected within the statutory exception for text and data mining for scientific research purposes (§ 60d UrhG) can be used to train AI models; under some approaches, the model itself, and at least some of its outputs, can be regarded as derived from the training data, which would fundamentally change their legal status. The discussion, closely monitored by ELSA, continues, and even more doubts were raised by the recent Getty Images lawsuit in the UK.

Finally, the use of generative AI raises important ethical questions, which also affect research data management. ELSA members proposed a taxonomy of ethical principles to be addressed at various stages of the research data lifecycle: Privacy, Property, Equality, Transparency and Freedom [3].

3. Discussion and Outlook

The section is open to the inclusion of further topics and willing to get an overview of newer questions such as the role of AI in RDM. A dialogue forum for the NFDI consortia is to be developed, which will serve to exchange legal, socio-scientific and research-ethical experiences and develop use-oriented, practical approaches to solutions. This is also explicitly aimed at the new consortia of the next NFDI approval rounds, for which the section is to offer a platform to constructively discuss overarching ethical and legal issues.

Competing interests

The authors declare that they have no competing interests.

Funding

The NFDI consortia where the authors are affiliated to are funded by the DFG.

References

1. Anna C. Haber, Ulrich Sax, Fabian Prasser, "Open tools for quantitative anonymization of tabular phenotype data: literature review", *Briefings in Bioinformatics*, Vol. 23, Issue 6, November 2022, bbac440, DOI: <https://doi.org/10.1093/bib/bbac440>.
2. Strathern et al., "QualiAnon - a Tool for anonymizing text data", Link: https://media-tum.ub.tum.de/doc/1575928/0hbkir8u5zldk4g2aw8icmnzo.TechReport_ws1.pdf.
3. Paweł Kamocki and Andreas Witt, "Ethical Issues in Language Resources and Language Technology – Tentative Categorisation", in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France. European Language Resources Association: 2022, pages 559–563.

Research Data Management Curriculum of the Research Data Services at the University Library Duisburg-Essen

Sophia Leimer¹[\[https://orcid.org/0000-0001-6272-204X\]](https://orcid.org/0000-0001-6272-204X), Sonja Hendriks¹[\[https://orcid.org/0000-0002-3460-4131\]](https://orcid.org/0000-0002-3460-4131), Lisa Korte¹[\[https://orcid.org/0000-0003-3650-2558\]](https://orcid.org/0000-0003-3650-2558), Jessica Stegemann¹[\[https://orcid.org/0000-0002-4149-1825\]](https://orcid.org/0000-0002-4149-1825), Sarah Ann Stock¹[\[https://orcid.org/0000-0003-4965-1085\]](https://orcid.org/0000-0003-4965-1085), Henning Timm¹[\[https://orcid.org/0000-0002-5345-7122\]](https://orcid.org/0000-0002-5345-7122), and Stephanie Rehwald¹[\[https://orcid.org/0000-0002-5884-4471\]](https://orcid.org/0000-0002-5884-4471)

¹ Research Data Services (RDS), University Library, University of Duisburg-Essen, Duisburg, Germany (email: rds.ub@uni-due.org, web: www.uni-due.de/rds)

Abstract. Reproducible and transparent science requires good research data management (RDM). However, researchers are often not familiar with how to appropriately handle research data, publication options, legal aspects, and the multitude of available software tools. To support scientists at the University of Duisburg-Essen in gaining RDM knowledge and competencies, the Research Data Services developed a structured teaching program – the RDM Curriculum – for researchers from different disciplines and career levels.

The RDM Curriculum is separated into a basic and an advanced module. In the basic module, participants can select from three full-day courses that address the basics of RDM with different disciplinary focus. The advanced module consists of various 2 hour online hands-on courses on diverse RDM topics and tools. As incentive to attend multiple courses, participants can earn the RDM Badge of the University Library Duisburg-Essen. The curriculum program is complemented by specialized RDM courses and self-learning materials.

With the RDM Curriculum, we created an easily accessible and individually selectable RDM course offer that enables researchers to gain discipline-specific and individually relevant RDM knowledge. The curriculum structure can gradually be expanded to further target groups (e.g. students) and allows for continuously adding new contents in the advanced module. The high request for the RDM Badge after the first iteration of the RDM Curriculum in the winter semester 2022/2023 confirmed the positive effect of incentives. Consequently, we are planning to establish a Data Champion Award as appreciation of RDM achievements and to showcase best-practice examples at the University of Duisburg-Essen.

Keywords: Training, Data Literacy, FAIR, Open Science

1. Introduction

Since practices of reproducible and transparent science are neither easy to apply nor self-explanatory, good training is indispensable for implementing good research data management (RDM). Researchers are often not familiar with how to appropriately handle research data, publication options, legal aspects, and the multitude of available software tools. At the same time, RDM methods and procedures vary between scientific disciplines. To support scientists at the University of Duisburg-Essen (UDE) in gaining RDM knowledge and competencies, the Research Data Services (RDS) developed a structured teaching program – the RDM Curriculum – for researchers from different disciplines and career levels. The Curriculum is

designed to teach general competencies in research data management as well as practical skills for RDM tools. In particular, RDM tools provided by the RDS are always accompanied by hands-on training to ensure a convenient and proper usage [1].

2. RDM Curriculum

2.1 Curriculum objectives

Following the university didactics method of constructive alignment [2], we first formulated curriculum objectives based on [3]. These represent the competencies and learning outcomes that participants will successively gain in the course program of the RDM Curriculum. We defined the following qualification goals:

- Participants know the different aspects of RDM along the research data life cycle.
- Participants are aware of local and national contact points (e.g. NFDI consortia) and consulting services (e.g. RDS).
- Participants are motivated to make their data FAIR and know methods, tools, and services to do so.
- In every-day research, participants can apply the different aspects of good RDM and handle individual subject-specific challenges.
- Participants actively exchange with other researchers about RDM.
- Participants can apply RDM methods and tools that are appropriate for their discipline and research.
- Participants have the ability to try out and apply new RDM methods.
- Participants can judge if an RDM tool or method is suitable for their research.


2.2 Structure of the RDM Curriculum

The RDM Curriculum is separated into a basic and an advanced module (Figure 1). In the basic module, participants can select from three courses that address the basics of RDM with different disciplinary focus: 1. (Bio-)Medicine with focus on electronic lab notebooks and sensitive data, 2. Education, Humanities, Social, and Economic Sciences with focus on sensitive data, text, audio, and video data, 3. STEM with focus on electronic lab notebooks, scripts, and code. These basic courses take place in person to facilitate exchange between participants. In 7 hours, an interactive introduction to RDM aspects along the research data life cycle, i.e. data management plans, metadata, data organization, storage, publication, and the respective focus topics, is given. The main target group are PhD students to establish basic RDM knowledge in the early scientific career. Other interested persons (e.g. Postdocs) are also welcome.


The advanced module consists of various 2 hour online courses on a diverse selection of specialized RDM topics and tools (e.g. RDM with Git and GitLab, Reproducible analyses with R Notebook). In these courses, we put high priority on introducing a specialized RDM topic or tool with hands-on and practical exercises and application examples. Accordingly, we expect active participation in course exercises (e.g. camera on, contributions). The target group for the advanced courses comprises PhD students, researchers, and interested persons (e.g. Master students, lab technicians) and does not require the prior participation in a basic course. Together with the relatively succinct online format, this should lead to an easily accessible and individually selectable course offer.

RDM Curriculum

Basic module	Advanced module	Special courses
<p>Basics RDM Subject recommendation: (Bio-) Medicine Focus: electronic lab notebooks, sensitive data</p>	<p>Coscine</p>	<p>Carpentries</p>
<p>Basics RDM Subject recommendation: Education, Humanities, Social and Economic Sciences Focus: sensitive data, text, audio & video data</p>	<p>Data reuse & reusable data</p>	<p>Introduction to R</p>
<p>Basics RDM Subject recommendation: STEM Focus: electronic lab notebooks, scripts & code</p>	<p>eLabFTW hands-on</p>	<p><i>in preparation: eLabFTW deep dive</i></p>
<p>Duration: 7 h Location: Duisburg or Essen Number of participants: min. 10, max. 20 Target group: PhD students Content: data management plans, metadata, data organisation and storage, publication Subject recommendation & focus: serves as orientation for course selection; courses freely selectable</p>	<p>RDM with Git & GitLab *</p>	<p>These events can be offered on request</p>
	<p>Project management with GitLab</p>	
	<p>Open Science Framework (OSF)</p>	
	<p>R Notebooks</p>	
	<p>Reproducible analysis workflows</p>	
	<p><i>in preparation: Dataverse</i></p>	
	<p>Duration: ca. 2 h Location: online (prerequisite: camera and microphone) Number of participants: min. 10, max. 20 Target group: PhD students, researchers and all interested persons</p>	
	<p>* Counted as two events</p>	
		<p>Other events</p>
		<p>Introducing the RDS</p>
		<p>RDS @ RDM for students</p>
		<p>RDS @ Research Academy Ruhr</p>
		<p>RDS @ Tag der Forschungsdaten</p>
		<p>RDS @ UDE Publication Days</p>
		<p><i>in preparation: RDS @ DataCampus</i></p>



UNIVERSITÄT
DUISBURG
ESSEN
Open-Minded



Research Data Services
Make your data

Figure 1. Overview of modules and courses of the Research Data Management (RDM) Curriculum.

After attending one course from the basic module and two courses from the advanced module, participants receive a certificate of attendance – the RDM Badge of the University Library Duisburg-Essen (Figure 2). The RDM Badge is intended to serve as incentive to deepen the RDM knowledge by participating in multiple courses and as demonstration of RDM training efforts.

The courses of the RDM Curriculum are usually offered once per semester for 10 to 20 participants per course. For information and course registration, we provide a comprehensive website (<https://www.uni-due.de/rds/en/rdmcurriculum.php>). Additionally, we offer specialized RDM courses on request (e.g. Data Carpentry, Introduction to R) and self-learning materials (e.g. RDM course on UDE's e-learning platform Moodle).



Figure 2. Research Data Management Badge of the University Library at the University of Duisburg-Essen (UDE), issued by the Research Data Services (RDS).

3. Conclusions and Outlook

The RDM Curriculum was offered the first time in the winter semester 2022/2023 and the second round of courses is currently ongoing. The number of course registrations was, as expected, lower in the initial phase, but increased with progressing advertisement. We especially experienced that promotion of the RDM Curriculum during the courses was very effective in increasing the number of registrations for upcoming courses. The group of participants ranged from PhD students from coordinated research programs to individual researchers from all career levels and also included Master students. Already at the end of the first round, we received several requests for the RDM Badge. This confirmed that the RDM Badge, as an incentive, indeed increased the willingness to learn about RDM. Consequently, we are planning to establish a Data Champion Award as appreciation of RDM achievements and to showcase best-practice examples at the University of Duisburg-Essen.

With the RDM Curriculum, we created an easily accessible and individually selectable RDM course offer that enables researchers to gain discipline-specific and individually relevant RDM knowledge and competencies. The structure of the RDM Curriculum can gradually be expanded to further target groups (e.g. Bachelor or Master students) and allows for continuously adding new contents in the advanced module.

Data availability statement

This work is conceptual and thus, there is no data necessary for replication.

Author contributions

SL: [Writing – original draft](#); SH, LK, JS, SAS, HT, SR: [Writing – review & editing](#); all authors: [Conceptualization](#); SL, JS: [Visualization](#); SR: [Supervision](#)

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

We thank the University Library Duisburg-Essen for their support, e.g. with course registration. Special thanks go to J. Pott for designing the RDM Badge. We are thankful to all people who gave constructive feedback during the curriculum development.

References

1. S. Rehwald and J. Stegemann, "Roadmap zur Servicestelle für Forschungsdatenmanagement am Beispiel der Universitätsbibliothek Duisburg-Essen," *Information - Wissenschaft & Praxis*, vol. 72, no. 4, pp. 194–203, Jul. 2021. doi: 10.1515/iwp-2021-2161.
2. J. Biggs and C. Tang, *Teaching for Quality Learning at University*, 4th ed. New York, NY, USA: Open University Press: 2011.
3. B. Petersen et al., "Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards," Sep. 2022, doi: 10.5281/ZENODO.7034478.

We are still here, too! Research Data Management at Universities of Applied Sciences

Approaches from the Project "FDM@HAW.rlp" in the German State Rhineland-Palatinate

Manuela Richter¹[\[https://orcid.org/0000-0003-1060-2622\]](https://orcid.org/0000-0003-1060-2622), Johannes Putzke²[\[https://orcid.org/0009-0006-8848-176X\]](https://orcid.org/0009-0006-8848-176X), Thomas M. Schimmer¹[\[https://orcid.org/0000-0003-0527-349X\]](https://orcid.org/0000-0003-0527-349X), and Anett Mehler-Bicher¹[\[https://orcid.org/0000-0003-1577-2276\]](https://orcid.org/0000-0003-1577-2276)

¹Mainz University of Applied Sciences, Germany

²Trier University of Applied Sciences, Germany

Abstract: The objective of the German non-profit association NFDI (German short form for "National Research Data Infrastructure") is to make the data stock of the entire German science system accessible to the public. To do so, it should involve all stakeholders. However, currently the Universities of Applied Sciences (UAS) are underrepresented in the NFDI, and there is a danger of neglecting their needs. Therefore, we present the project "Research Data Management at Universities of Applied Sciences in the State of Rhineland-Palatinate" (FDM@HAW.rlp), which is funded by the German Federal Ministry of Education and Research (BMBF) and financed within the Recovery and Resilience Facility of the European Union. In the project, seven public UAS in Rhineland-Palatinate and the Catholic University of Applied Sciences (CUAS) Mainz follow a common goal: They intend to establish an institutional RDM within a period of three years by building up competencies at the UAS, setting up services for researchers and finding solutions for a common technical infrastructure.

Keywords: institutional RDM, Universities of Applied Sciences, FDM@HAW.rlp, spreading RDM

1 Extended Abstract

The objective of the German non-profit association NFDI (German short form for "National Research Data Infrastructure") is to make the data stock of the entire German science system accessible according to the FAIR principles [1], [2]. This requires not only technological and methodological solutions, but also communication and networking among various stakeholders.

In networking, the NFDI encounters two challenges: First, linking the already existing initiatives (especially initiatives at the state level). Second, connecting the Universities of Applied Sciences (UAS) more closely with the NFDI. Currently, the UAS are severely underrepresented in the NFDI. Rather, the NFDI consortia are dominated and shaped by the universities (one of few exceptions is Mainz University of Applied Sciences,

which is a co-applicant in the NFDI4Objects consortium since March 2023). In principle, the mission and focus of universities is basic research, while the UAS conduct application-oriented research. At the same time, it must be noted that UAS were primarily focused on teaching in the past. For several years now, however, they have also been conducting increasingly successful research and so the needs in the field of RDM are also growing steadily.

For the efficient development of a national infrastructure, it is essential to involve all relevant stakeholders, to create synergies and to integrate what already exists. In a position paper [3], the state initiatives have already shown that their continuation and inclusion in the NFDI is a prerequisite for success.

However, the development of RDM competencies and structures at the level of single UAS requires further efforts. This is due to the structure, size, and orientation of UAS in Germany. Compared to universities, only very few UAS have the necessary infrastructures such as large computing centers and libraries. In terms of human resources, the difference between UAS and universities is due to the lower level of funding, the lack of mid-level academic staff and the increased teaching load of the scholars in the UAS. Research projects at UAS are more often carried out in cooperation with industry and small and medium enterprises, so that aspects of data protection and data governance play an important role. An essential building block for the establishment of RDM is the education and training of all stakeholders in data literacy.

The German Federal Ministry of Education and Research (BMBF) has recognized these general conditions and has announced the funding guideline "Reuse and Management of Research Data at Universities of Applied Sciences" [4]. The aim is to identify the needs of UAS, raise awareness of RDM, and establish and expand RDM. Funding is being provided for 14 projects for structural development and three studies to determine the status quo at UAS [5].

The preliminary results of the study "Development and Dissemination of Research Data Management at Universities of Applied Sciences" (EVER.FDM, [6]) show that the topic is still in its initial stages. The guideline on Good Scientific Practice [7] by the German Research Foundation (DFG) was found to be known to 79% of respondents, but the FAIR principles were found to be known to only 25.3%. With only 9.8% awareness, the Stifterverband's Data Literacy Charta [8] scored even worse. This shows that there is a need for essential work at the basis, especially in the area of data literacy, the creation of structures, and networking with the NFDI.

These challenges are met by the project "Research Data Management at Universities of Applied Sciences in the State of Rhineland-Palatinate" (FDM@HAW.rlp). In the project, the seven state UAS in Rhineland-Palatinate and the CUAS (see Figure 1) follow a common goal: They aim to establish an institutional RDM within a period of three years by building up competencies, services and technical infrastructure.

The project is organized as follows: At each UAS, there is located a "scout" who acts as a primary contact person for the researchers at the respective institution. Furthermore, there are 4 "data stewards" who are responsible for a scientific domain across institutions. These domains are: (1) life sciences and natural sciences, (2) informatics / computer science, (3) engineering, and (4) humanities and social sciences.

Both are also responsible for educational offerings that address the fundamentals of RDM. The stewards are also responsible for networking with the NFDI to provide researchers with tailored solutions to their RDM concerns. The project is coordinated by a central coordination unit. After the duration of the project, the positions of "scouts",

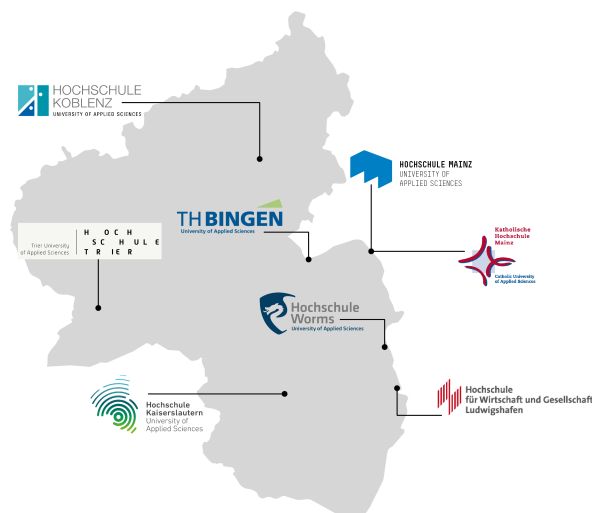


Figure 1. Participating Universities of Applied Sciences / Illustration: FDM@HAW.rlp.

"data stewards" as well as "central coordination unit" are supposed to be merged into a single RDM officer position at each of the eight universities.

In the presentation, the authors will tell the inside story regarding the tasks and experiences of Scouts and Stewards in everyday work. In doing so, they focus on the specific needs and requirements of RDM at UAS in order to actively bring them into the discussion and to include UAS with their challenges and needs in the NFDI's deliberations. Only when the entire science system in Germany is networked and involved, the NFDI can be successful on a broad scale. It is therefore a necessity that the various projects network with each other. Likewise, the needs of the UAS must be formulated and the NFDI's offerings must be made available to them. An important building block for spreading RDM is the development of data literacy of all those involved.

Author contributions

Manuela Richter: Writing - original draft

Dr. Johannes Putzke: Writing - review & editing

Dr. Thomas M. Schimmer: Writing - review & editing, Project administration

Prof. Dr. Anett Mehler-Bicher: Project administration

The authors hereby declare that there are no competing interests.

Funding

This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within the funding measure Reuse and Management of Research Data at Universities of Applied Sciences (funding number 16FDFH104A) and financed within the Recovery and Resilience Facility of the European Union. The authors are responsible for the content of this publication.



SPONSORED BY THE

Federal Ministry
of Education
and Research



Funded by
the European Union

NextGenerationEU

Acknowledgements

The authors want to acknowledge the partner institutions of the project: Mainz University of Applied Sciences (funding number 16FDFH104A), Bingen Technical University of Applied Sciences (funding number 16FDFH104B), Hochschule Kaiserslautern University of Applied Sciences (funding number 16FDFH104C), Koblenz University of Applied Sciences (funding number 16FDFH104D), Ludwigshafen University of Business and Society University of Applied Sciences (funding number 16FDFH104E), Trier University of Applied Sciences (funding number 16FDFH104F), Hochschule Worms University of Applied Sciences (funding number 16FDFH104G), Catholic University of Applied Sciences Mainz (funding number 16FDFH104H).

References

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship, journal = Scientific Data," vol. 3, no. 1, Mar. 2016. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18). [Online]. Available: <https://doi.org/10.1038/sdata.2016.18>.
- [2] National Research Data Infrastructure (NFDI) e.V. "The German National Research Data Infrastructure - Association." (2020), [Online]. Available: <https://www.nfdi.de/association/?lang=en> (visited on 04/20/2023).
- [3] A. Axtmann, E. Böker, O. Brand, *et al.* "Wir bringen die breite Basis mit – Gemeinsames Plädoyer für eine enge Einbindung der Landesinitiativen für Forschungsdatenmanagement in die Nationale Forschungsdateninfrastruktur." (2021), [Online]. Available: <https://zenodo.org/record/4524655>.
- [4] Federal Ministry of Education and Research. "Bekanntmachung - Richtlinie zur Förderung von Projekten zum Thema Nachnutzung und Management von Forschungsdaten an Fachhochschulen, (Bundesanzeiger vom 17.08.2021)." (2021), [Online]. Available: <https://www.bmbf.de/bmbf/shareddocs/bekanntmachungen/de/2021/08/2021-08-17-Bekanntmachung-Fachhochschulen.html> (visited on 04/20/2023).
- [5] Federal Ministry of Education and Research. "Forschungsdatenmanagement - Förderung von 14 Projekten gestartet." (2021), [Online]. Available: https://www.bildung-forschung.digital/digitalezukunft/de/wissen/Datenkompetenzen/forschungsdatenmanagement_fachhochschulen/forschungsdatenmanagement_fachhochschulen_node.html (visited on 04/20/2023).
- [6] R. Werth. "Entwicklung und Verbreitung von FDM an Fachhochschulen – Eine bundesweite empirische Analyse zu Aktivitäten und Bedarfen, Vortrag RDA Deutschland Tagung."

- (2023), [Online]. Available: https://indico.desy.de/event/37011/contributions/132893/attachments/80170/104661/2023-02-14_EVER_FDM%40RDA-DE_v1.1.pdf (visited on 04/20/2023).
- [7] German Research Foundation. "Guidelines for Safeguarding Good Research Practice - Code of Conduct." (2022), [Online]. Available: https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp_en.pdf (visited on 04/20/2023).
- [8] K. Schüller, H. Koch, and F. Rampelt. "Data-Literacy-Charta. Version 1.2. Berlin: Stifterverband." (2021), [Online]. Available: https://www.stifterverband.org/sites/default/files/data-literacy-charta_v1_2.pdf (visited on 04/20/2023).

FAIR and scalable education

The Galaxy training network (GTN) and a Training Infrastructure as a Service (TlaaS)

Anika Erxleben-Eggenhofer¹[\[https://orcid.org/0000-0002-7427-6478\]](https://orcid.org/0000-0002-7427-6478), Teresa Müller²[\[https://orcid.org/0000-0003-1252-9684\]](https://orcid.org/0000-0003-1252-9684), and Bérénice Batut¹[\[https://orcid.org/0000-0001-9852-1987\]](https://orcid.org/0000-0001-9852-1987)

¹ Freiburg Galaxy Team, Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany

² Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg, Germany

Abstract. The Galaxy Project [1][2] is a widely-used open-source platform for data-driven research that offers an extensive suite of tools and services for analyzing and visualizing large-scale data. Besides Galaxy as data analysis framework, the Galaxy Training Network (GTN; [3][4]) provides a huge collection of training material for researchers, developers and trainers. Among the various features offered by the Galaxy platform, the GTN and the Training Infrastructure as a Service (TlaaS; [5]) stands out as an innovative solution for delivering scalable and flexible training to researchers and educators. TlaaS is built on top of the Galaxy platforms which enable users to request dedicated compute resources for training purposes. This infrastructure, in combination with GTN, provides an excellent solution for delivering high-quality training to researchers and educators around the world, regardless of their geographical location or hardware limitations. One of the key benefits of the GTN is its ability to enable users to create and share training materials for a variety of data analysis workflows, including e.g. genomics, proteomics, and metabolomics but also for imaging and climate data. These training materials can be delivered in various formats, including interactive tutorials, videos, and webinars, and can be accessed from any location with an internet connection. Moreover, the platform offers a wide range of community-contributed training materials that cover a vast array of topics in biomedical research and beyond, making it an invaluable resource for researchers and educators. To use TlaaS, instructors simply need to register their training event in a simple form to get access to the dedicated compute resources.

TlaaS was created by the implementation of two components: a web service, and a default set of Galaxy job scheduling rules, which function together to present a private queue for users in specific Galaxy user groups. The web service enables registering requests for resources for the training event specifying time, topic with tools to be used and number of trainees. Additionally then a training group is created in Galaxy by adding members to those groups in a GDPR compliant way, as needed. TlaaS coordinators or system administrators review these requests, using information about the class size, the tools used in the training materials, as well as the resource allocations of those tools on the infrastructure, to estimate the required compute resources.

Training participants access a specific training URL at the start of the training event, after which they are automatically registered in the TlaaS system without further user interaction and without instructors needing to manually manage group membership. The job scheduler, once aware of the training group, will place any job run by someone in that training onto the dedicated training nodes (Fig. 1).

In this talk we will highlight how Galaxy and TlaaS in combination with well curated training material can scale to thousands of simultaneous trainees and enable Massive Open Online Course (MOOC) events that we have been executing the last 3 years during the pandemic.

Keywords: Galaxy, Galaxy Training Network, training, education, TlaaS

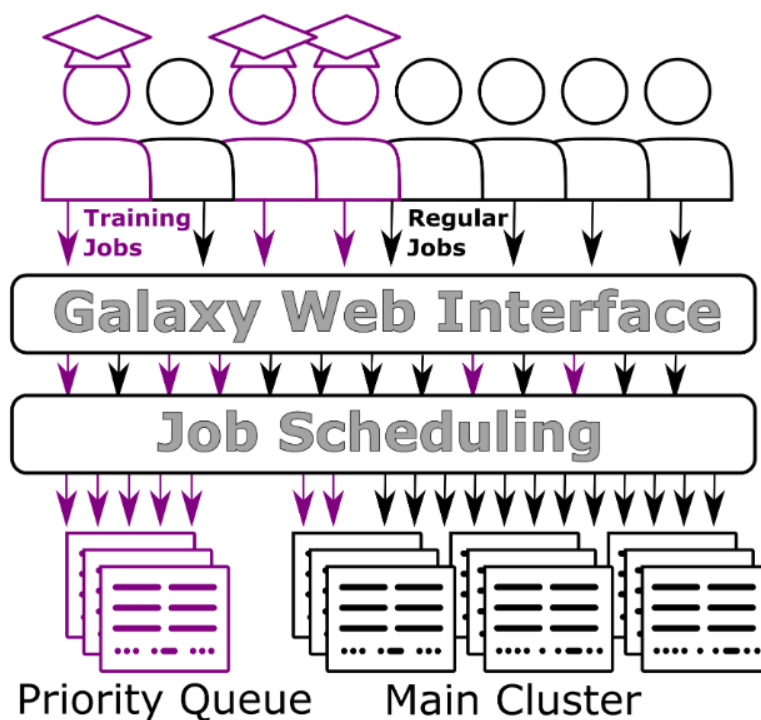


Figure 1. Scheme of the TlaaS queuing system. Jobs are processed by the same Galaxy server, but when those jobs come from users in the training group, they receive special handling.

Competing interests

The authors declare that they have no competing interests.

References

1. The Galaxy Community - Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update; Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351, <https://doi.org/10.1093/nar/gkac247>
2. <https://galaxyproject.org> (accessed Apr. 24, 2023).
3. Hiltmann & Rasche et al. - Galaxy Training: A powerful framework for teaching! PLoS Comput Biol 19(1): e1010752. <https://doi.org/10.1371/journal.pcbi.1010752>
4. <https://training.galaxyproject.org> (accessed Apr. 24, 2023).
5. <https://usegalaxy-eu.github.io/tiaas.html> (accessed Apr. 24, 2023).

Der Zertifikatskurs „Forschungsdatenmanagement“ als Blaupause für die FDM-bezogene Kompetenzentwicklung im Rahmen der NFDI

Benjamin Slowig¹[\[https://orcid.org/0000-0001-5343-2788\]](https://orcid.org/0000-0001-5343-2788), Mirjam Blümm²[\[https://orcid.org/0000-0003-3665-7031\]](https://orcid.org/0000-0003-3665-7031),
Konrad U. Förstner³[\[https://orcid.org/0000-0002-1481-2996\]](https://orcid.org/0000-0002-1481-2996), Birte Lindstädt⁴[\[https://orcid.org/0000-0002-8251-1597\]](https://orcid.org/0000-0002-8251-1597),
Rabea Müller⁵[\[https://orcid.org/0000-0002-3096-8237\]](https://orcid.org/0000-0002-3096-8237), and Marvin Lanczek⁶[\[https://orcid.org/0000-0001-8247-3052\]](https://orcid.org/0000-0001-8247-3052)

¹ Landesinitiative fdm.nrw, Universität Duisburg-Essen

² Technische Hochschule Köln

³ ZB MED – Informationszentrum Lebenswissenschaften / Technische Hochschule Köln

⁴ ZB MED Informationszentrum Lebenswissenschaften

⁵ ZB MED Informationszentrum Lebenswissenschaften

⁶ ZBIW - Technische Hochschule Köln

Abstract. Der Zertifikatskurs Forschungsdatenmanagement hat sich als erstes strukturiertes und berufsbegleitendes Weiterbildungsangebot im Bereich des FDM-Kompetenzaufbau etabliert. Als Grundlage für eine Weiterentwicklung in der NFDI rückt vor allem dessen disziplinspezifische Adaption durch die Fachkonsortien in den Fokus. Dies geschieht u. a. im Rahmen der Sektion EduTrain. Neben der konzeptionellen Weiterentwicklung spielt das Ausrollen des Zertifikatskurses über die NFDI oder andere Akteure, wie die Landesinitiativen, aktuell eine große Rolle. Hiermit soll dem nachgewiesenen Bedarf an qualifizierten Personal im Forschungsdatenmanagement gemeinsam begegnet werden.

Keywords: Weiterbildung, Zertifikatskurs, Forschungsdatenmanagement, Sektion EduTrain, NFDI

1. Der Zertifikatskurs Forschungsdatenmanagement

Der Zertifikatskurs „Forschungsdatenmanagement“ stellt das erste berufs begleitende, strukturierte und zertifizierte Qualifizierungsangebot zum FDM in Deutschland dar, welches auf Basis einer Kooperation zwischen der TH Köln, dem dort ansässigen ZBIW, ZB MED und der Landesinitiative fdm.nrw initiiert und fortlaufend ausgerichtet wird. Zielgruppen sind hierbei Beschäftigte der forschungsnahen Infrastruktureinrichtungen (insb. Bibliotheken, Rechenzentren und Forschungsförderung), zentral agierende (generic) Data Stewards sowie dezentral bzw. in den Fach-/Forschungsbereichen angesiedelte (embedded) Data Stewards. Innerhalb der Module werden den Teilnehmenden fundierte Kenntnisse und Fähigkeiten im Umgang mit Forschungsdaten vermittelt. Damit werden die Teilnehmenden dazu befähigt, Forschungsdaten während des gesamten Forschungslebenszyklus, von der Planung über die Erhebung und Speicherung bis zur Archivierung und Wiederverwendung, professionell zu managen. Der Kurs beinhaltet eine Vielzahl von Aspekten, wie Datenmanagementpläne, Metadatenstandards, rechtliche Aspekte des Datenmanagements, Qualitätssicherung von Daten und vieles mehr.

Auf Basis der Erfahrungen aus den ersten Durchgängen des Zertifikatskurses (exklusiv für NRW), der Vernetzungs- und Austauschformate der FDM-Community, der hohen Anzahl an Stellenausschreibungen für den Bereich FDM und die NFDI sowie der aktuellen Anmeldephase für den dritten Durchgang (für den Bewerbungen aus ganz Deutschland eingereicht werden konnten) ist ein wachsender Bedarf an solch einem Qualifizierungsangebot zu verzeichnen. Eine Erhöhung der Ausbildungskapazitäten durch ein Ausrollen des Konzepts ist daher wünschenswert und bereits in der Planung. Inhaltlich wird der Zertifikatskurs „Forschungsdatenmanagement“ von Seiten der NFDI-Konsortien und weiteren Initiativen und Standorten als anerkanntes Angebot für den Kompetenz-Aufbau und -Ausbau verschiedener Zielgruppen gesehen. Insbesondere im Rahmen der NFDI-Sektion „Training and Education“ (EduTrain) wird intensiv über die Struktur, die Inhalte und Lernziele diskutiert und inwieweit sich dieser Kurs als Vorlage für weitere Formate (bzw. Module) zur Kompetenzvermittlung insbesondere für die Zielgruppe der Data Stewards (generic und embedded) eignet.

1.1 Vorgehen

Aus diesen Diskussionen heraus ist ein Antragsvorhaben im Rahmen von Base4NFDI entstanden. Das Ziel besteht darin, den Zertifikatskurs als attraktiven Basis-Service für die Konsortien der Nationalen Forschungsdateninfrastruktur (NFDI) zu öffnen und gleichzeitig eine Blaupause für die Entwicklung weiterer disziplinspezifischer FDM-Module (ggf. im Rahmen weiterer Ausbauphasen in Base4NFDI) zu schaffen. Durch das fortlaufende Angebot und die angedachte Erweiterung des Kurses werden immer mehr embedded Data Stewards der verschiedenen Fachbereiche qualifiziert und parallel Ziele der einzelnen Konsortien (und deren Arbeitspakete zum Aspekt Training) gemeinsam verfolgt.

Die Ausgestaltung des Antrags erfolgt in enger Zusammenarbeit mit den NFDI-Konsortien, um die Inhalte und Lernziele fachspezifisch und passgenau zu gestalten. Grundlage dafür bildet die von der Unter-AG Schulungen/Fortbildungen (DINI/nestor, AG Forschungsdaten) publizierte und generisch ausgerichtete Lernzielmatrix zum FDM, die von den Konsortien fachspezifisch auf die embedded Data Stewards übertragen wird. Der Erfahrungsaustausch und die Vernetzung mit allen NFDI-Konsortien findet im Rahmen der Sektion "Training and Education" (EduTrain) statt. Auf diese Weise soll nicht nur die Anschlussfähigkeit an das Projekt „Data Literacy Alliance“ (DALIA) gewährleistet werden, sondern insgesamt das Adaptionspotential gesteigert werden. Es ist zudem angedacht, Inhalte und Materialien zur freien Nachnutzung zur Verfügung zu stellen, wodurch Vorlagen für weitere Module bzw. Qualifizierungsangebote geschaffen werden.

1.2 Weiterführende Fragen

Auf der Basis der Informationen zu den aktuellen Prozessen zum Zertifikatskurs, der Zusammenarbeit im Rahmen der NFDI sowie zum angesprochenen Antrag stellen sich folgende weiterführende Fragen, die diskutiert werden sollten:

- In welchen Bereichen bzw. bei welchen Themen gibt es Bedarf, die Inhalte und Module des Zertifikatskurses zu erweitern?
- Wie müsste der Kurs gestaltet sein, um die Lernziele und Bedarfe von fachspezifisch arbeitenden Data Stewards abzudecken?
- Was ist der Mehrwert von Selbstlerneinheiten (z. B. als OER) zum FDM für den Kompetenz-Aufbau und -Ausbau und worin liegen deren Grenzen?

Data availability statement

nicht zutreffend

Underlying and related material

Modulhandbuch Zertifikatskurs Forschungsdatenmanagement 2023/24: https://www.th-koeln.de/mam/downloads/deutsch/weiterbildung/zbiw/angebote/zbiw_modulhandbuch_zk_fdm_2023-24.pdf (pdf, 1024 KB)

Author contributions

Benjamin Slowig: Writing – original draft

Mirjam Blümm: Writing – review & editing

Konrad U. Förstner: Conceptualization

Birte Lindstädt: Writing – review & editing

Rabea Müller: Writing – review & editing

Marvin Lanczek: Conceptualization

Competing interests

nicht zutreffend

Funding

nicht zutreffend

Acknowledgement

Für den Zertifikatskurs „Forschungsdatenmanagement“ werden Stipendien über die Digitalisierungsoffensive des Ministeriums für Kultur und Wissenschaft NRW für förderfähige Institutionen vergeben, die einschließlich des 2023 startenden Durchgangs zur Verfügung stehen.

References

1. “Zertifikatskurs Forschungsdatenmanagement.” TH Köln. https://www.th-koeln.de/weiterbildung/zertifikatskurs-forschungsdatenmanagement_82048.ph (accessed Apr. 26 2023).
2. B. Slowig et al., “Der Zertifikatskurs Forschungsdatenmanagement in NRW: Eine modular aufgebaute Weiterqualifikation für das professionelle Datenmanagement,” *O-Bib. Das Offene Bibliotheksjournal*, vol. 9, no. 3, pp. 1-10, Aug. 2022, doi: 10.5282/o-bib/5833.
3. M. Blümm et al., “Der Zertifikatskurs Forschungsdatenmanagement als adaptierbares Aus- und Weiterbildungsangebot,” in: *E-Science-Tage 2021: Share Your Research Data*, V. Heuveline and N., Eds. Apr. 2022, pp. 414–420, doi: 10.11588/heibooks.979.c13758.
4. B. Petersen et al., “Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards,” Zenodo, Sep. 5, 2022, doi: 10.5281/zenodo.7034478.

Building Research Data Management (RDM) expertise and training resources in ELIXIR Nodes

Celia van Gelder¹[\[https://orcid.org/0000-0002-0223-2329\]](https://orcid.org/0000-0002-0223-2329), Alexia Cardona², Brane Leskosek³, Patricia Palagi⁴

¹ ELIXIR-NL

² ELIXIR-UK

³ ELIXIR-SI

⁴ ELIXIR-CH

Research data Management (RDM) is central to the implementation of the FAIR (Findable Accessible, Interoperable, Reusable) and Open Science principles. Recognising the importance of RDM, the ELIXIR-CONVERGE project was launched in 2020 and included a work package on Training & Capacity Building in Data Stewardship, in which 21 Nodes participate. Since there is a clear need for dissemination of RDM know-how, practices and resources and a demand for RDM training material and courses, for researchers, trainers and RDM professionals, the aim of the work package was to build a comprehensive, interconnected and sustainable RDM Training Portfolio across all Nodes, and to increase the RDM training expertise and capacity in ELIXIR Nodes.

RDM professionals, training experts and trainers from ELIXIR and beyond identified gaps in the Nodes' RDM training programmes and defined priority topics (DMP, Data Stewardship, FAIR/metadata and Reproducibility) (Cardona et al, 2021) for training development. Collaborative work from all the ELIXIR Nodes during ELIXIR-CONVERGE resulted in an extensive Data Management/Data Stewardship (DM/DS) course portfolio (Cardona et al, 2021)¹ which aligns with the Nodes' RDM training strategies. The portfolio includes both generic DM/DS resources and materials, as well as specific materials related to the priority topics mentioned above, and it also includes training materials for specific domains (e.g. the Plant Demonstrator). The course portfolio is available in [TeSS](#), in which training materials are tagged with RDM terms to increase their findability, and links to relevant training materials were included in [RDMkit](#). In collaboration with the ELIXIR Training Platform, the learning paths methodology has been applied to RDM for data stewards and researchers (e.g. with the Plants Sciences and the System Biology Communities), and work is ongoing to extend the [ELIXIR-GOBLET Train-the-Trainer programme](#) to tackle the challenges that exist in teaching RDM. Moreover, during ELIXIR-CONVERGE, a strong RDM trainers network was formed, establishing solid RDM training expertise within the Nodes. This emerging network raised great interest in the Nodes, showing their willingness to work together on developing training materials, sharing their knowledge, and establishing best practices and standards for RDM training.

The main RDM training related actions in ELIXIR going forward are: (i) gathering the RDM courses in the portfolio into learning paths for specific audiences, thus shaping the

¹ Updated numbers (February 2020 - December 2022): 18 Nodes contributed to 119 training events with 2849 participants.

ELIXIR RDM curriculum, (ii) increasing findability and reusability of the developed courses and modules by optimising annotation and linking in TeSS, RDMkit and FAIR Cookbook, (iii) encouraging (re)use of the RDM materials in ELIXIR Nodes, and (iv) increasing RDM training capacity in Nodes by finalising and launching the RDM Train the Trainer programme.

In 2023, the ELIXIR RDM Community was initiated, and the network of RDM trainers will be an integral part of this Community. This will enable the upskilling of RDM trainers and scaling up RDM training capacity in the Nodes, while bridging to RDM professionals outside ELIXIR as well.

References

1. Cardona, Alexia, van Gelder, Celia, Palagi, Patrícia, & Leskošek, Brane. (2021). ELIXIR-CONVERGE D2.1 Report on the identified training needs and a timeline for training courses. Zenodo. <https://doi.org/10.5281/zenodo.4892863>

RDM in Chemistry: How to Educate and Train Future Researchers to Manage Their Data

Jochen Ortmeyer¹[\[https://orcid.org/0000-0003-2074-8027\]](https://orcid.org/0000-0003-2074-8027), Fabian Fink¹[\[https://orcid.org/0000-0002-1863-2087\]](https://orcid.org/0000-0002-1863-2087),
Alexander Hoffmann¹[\[https://orcid.org/0000-0002-9647-8839\]](https://orcid.org/0000-0002-9647-8839), and Sonja Herres-Pawlis¹[\[https://orcid.org/0000-0002-4354-4353\]](https://orcid.org/0000-0002-4354-4353)

¹ RWTH Aachen University, Germany

Abstract. For in-depth research data management in chemistry, a cultural change is inevitable. To foster this change, future researchers need to be educated accordingly. The presentation will provide an overview of the first teaching approaches in student courses in chemistry at RWTH Aachen University. On the long range, the integration into curricular teaching is key to the cultural change.

Keywords: Chemistry, Education, Training, Data Management, Curricular Teaching

1. Introduction

More and more digital research data are generated in Chemistry. So, new concepts are essential: In which data formats can data be stored in the long term? How and where can data be stored? Which information of the experiment / the simulation should be noted in the metadata? How can these data be made accessible for group members and other researchers? How can these data be made findable for researchers and AI algorithms?

2. Teaching Approach

Researchers need to be trained in these topics and concepts to apply them successfully in their daily research processes. NFDI4Chem[1-4] tackles these challenges by providing several teaching and training courses and materials – for all career stages as well as all RDM levels (**Figure 1**). For example, we offer regularly different workshops on research data management or electronic lab notebooks. Furthermore, we believe that young chemists and students are key to the cultural change. Therefore, we are increasingly paying attention on education, e.g., providing teaching courses, teaching materials, and knowledge bases. This presentation will highlight the first teaching programs implemented in theoretical and practical student courses and collected feedback of students thereof. We have collected several years of experience on the integration of RDM into an inorganic student's lab course in the fifth semester, as a hands-on experience where students prepare special chemical substances and directly use the electronic lab notebook Chemotion to document this. Further, we deepen the RDM discussion in a Master lecture on sustainable polymerisation catalysis with case studies on good and bad RDM. Here, best practices from NFDI4Chem directly serve as case studies and are used for teaching. The students go into their Master theses having already the basics of RDM in mind which is an excellent prerequisite to apply on their own research data later on. The integration of these RDM units into the existing lab course and the existing Master lecture did not require large curricular changes which would normally need years at universities. In our approach, we

integrated RDM units seamlessly into the existing study programme which allows a fast integration into the actually running chemistry programmes. This method may serve as best practice also for other disciplines and their efforts to integrate RDM teaching into their study programmes.

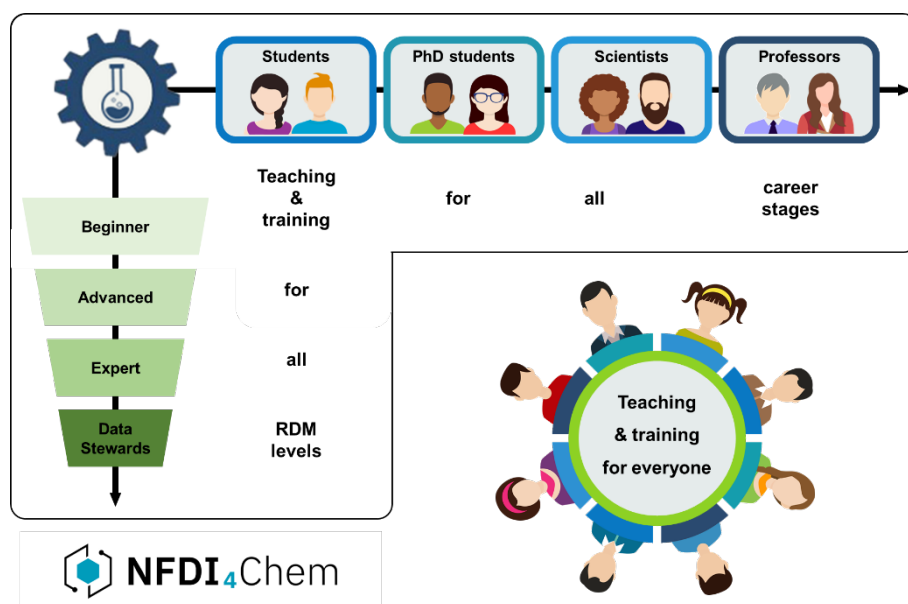


Figure 1. Multidimensional teaching and training for the chemistry community.

Data availability statement

Does not apply in total. Data of collected feedback to be published until presentation.

Underlying and related material

Does not apply.

Author contributions

Jochen Ortmeyer: Writing, Visualization, Presentation

Fabian Fink: Formal Analysis, Visualization

Alexander Hoffmann: Supervision, Conceptualization

Sonja Herres-Pawlis: Project administration, Supervision, Conceptualization

Competing interests

The authors declare that they have no competing interests.

Funding

NFDI4Chem is supported by DFG under project number 441958208.

Acknowledgement

The authors gratefully acknowledge support by all members of NFDI4Chem.

References

1. S. Herres-Pawlis, O. Koepler, C. Steinbeck, „NFDI4Chem: Shaping a Digital and Cultural Change in Chemistry”, *Angew. Chem. Int. Ed.*, 58, 32, 10766-10768, August 2019, doi: <https://doi.org/10.1002/anie.201907260>
2. K. C. Steinbeck, O. Koepler et al., “NFDI4Chem – Towards a National Research Data Infrastructure for Chemistry in Germany”, *Research Ideas and Outcomes*, 6, e55852.2020, June 2020, doi: <https://doi.org/10.3897/rio.6.e55852>
3. J. Ortmeyer, J. D. Jolliffe, „Treatment of research data“, *Nachr. Chem.*, 70, 10, 16-17, Oktober 2022, doi: <https://doi.org/10.1002/nadc.20224131398>
4. NFDI4Chem. “NFDI4Chem”. NFDI4Chem. <https://www.nfdi4chem.de/> (date accessed)

Experiences from FAIRifying community data and FAIR infrastructure in biomedical research domains

Dagmar Waltemath¹[\[https://orcid.org/0000-0002-5886-5563\]](https://orcid.org/0000-0002-5886-5563), Esther Thea Inau¹[\[https://orcid.org/0000-0002-8950-2239\]](https://orcid.org/0000-0002-8950-2239),
Lea Michaelis¹[\[https://orcid.org/0000-0001-9691-2677\]](https://orcid.org/0000-0001-9691-2677), Venkata Satagopam²[\[https://orcid.org/0000-0002-6532-5880\]](https://orcid.org/0000-0002-6532-5880), and
Irina Balaur²[\[https://orcid.org/0000-0002-3671-895X\]](https://orcid.org/0000-0002-3671-895X)

¹ Medical Informatics Laboratory and Core Unit Data Integration Center, University Medicine Greifswald, Greifswald, Germany

² Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg

Abstract. FAIR data is considered good data. However, it can be difficult to quantify data FAIRness objectively, without appropriate tooling. To address this issue, FAIR metrics were developed in the early days of the FAIR era. However, to be truly informative, these metrics must be carefully interpreted in the context of a specific domain, and sometimes even of a project. Here, we share our experience with FAIR assessments and FAIRification processes in the biomedical domain. We aim to raise the awareness that “being FAIR” is not an easy goal, neither the principles are easily implemented. FAIR goes far beyond technical implementations: it requires time, expertise, communication and a shift in mindset.

Keywords: FAIR assessment, FAIRification, biomedicine, EOSC, COMBINE

1. Introduction

The Findability, Accessibility, Interoperability and Reusability (FAIR) guiding principles for scientific data management and stewardship [1] provide guidance for sustainable deposition and sharing of scientific outcomes. Adherence to these principles leads to a more systematic approach for machine-actionable work with scientific data. FAIR data has a better chance to be reproducible, is more trustworthy and is cited more often. Large research networks foster the exchange and discussion about FAIR data, leading to new concepts, methods, scientific infrastructures and tools.

FAIR-related actions, however, need to be consolidated to understand what communities mean by “FAIR”, to ensure communication about reusable Research Data Management tools, to foster cross-community developments and to build common metadata standards. These needs have already been identified. For example, the Research Data Alliance [2] developed the FAIR Data Maturity Model [3], a standard reference system for FAIR assessment tools and FAIRification workflows.

Our experiences with FAIR evaluations and FAIRification tasks show that scientists overestimate the FAIRness of their data rather than being too critical. Oftentimes the FAIR scores are a surprise and even a disappointment. We also observe that while FAIR principles are generally known, only a few scientists can explain their meaning or interpretation in detail. Thus, it is imperative to educate scientists on what it actually means to strive for a “FAIRness” and to support them in the FAIRification process.

2. Results

We assessed five infrastructures and performed subsequent FAIRification.

1. The **Computational Modeling in Biology Network** [4] coordinates the development of community standards and formats for computational models in biomedicine. Building on the experience of mature projects, COMBINE continuously integrates new requirements for model sharing and reuse, e.g. harmonised metadata [5], model-specific FAIR metrics (<https://github.com/FAIR-CA-indicators>), and guidelines for FAIR data sharing [6]. Funded by the EOSC Future, we develop a domain-specific FAIR assessment tool. This community-oriented process allows us to work with cross-domain experts.
2. The **Disease Maps Project** (<https://disease-maps.org>) develops disease-specific comprehensive knowledge representations. The development of a disease map is a complex process requesting participants with interdisciplinary expertise over a considerable time period [7]. Therefore, it is essential that the maps are available to a broad scientific community in order to benefit from the invested efforts. We assessed the FAIRness of the COVID-19 Disease Map in the Molecular Interaction NETwoRK Visualization (MINERVA) platform [8] following the template provided in the IMI FAIRplus Project (<https://fairplus-project.eu/>). MINERVA is a FAIR infrastructure, facilitating the discovery and the accessibility of the integrated biological content, supporting the authorisation/ authentication features, providing a licensing system at diagram/ project level and integrating a converter for systems biology standards, thus supporting interoperability.
3. The **Study of Health in Pomerania** is a population-based cohort study designed to investigate the long-term progression of subclinical findings, their determinants and prognostic values [9]. We specifically explored the FAIRification of the medical laboratory metadata, leading to the indication that successful FAIRification requires interdisciplinary collaboration between data stewards and domain experts.
4. The **German Center for Diabetes Research** has established a Core Dataset for diabetes research (<https://medical-data-models.org/45430>). Its FAIRification improves reusability and study comparability. Hence, we structured the data, mapped the codes to terminologies, and implemented a formalised, provenance-enabled and semantically enriched representation of (meta)data. A baseline FAIRness evaluation helped us plan the FAIRification and establish a fruitful collaboration between the data owners, clinicians and data curators.
5. The **German Network University Medicine** supports the COVID-19 data collection from German University Hospitals. Adherence to the FAIR principles has been discussed from the start of the project. In 2023, we set out to assess the FAIRness level of the overall project, its sub-projects and domain-specific datasets using a manual evaluation system [10]. Interestingly, the participants reported only little knowledge of the FAIR principles, and questions addressing the uptake of FAIR recommendations showed that knowledge about the actual data management processes had been missing.

3. Discussion and lessons learnt

Working with the FAIR principles is challenging and the FAIR journey of a research institution requires actions, change of workflows and mindsets and financial support. Data owners, data managers, scientists, stakeholders and funding agencies need to actively contribute at each step of the data lifecycle, from design/ collection to sustainability, to FAIRify data, to minimise errors, and ultimately, to save time and to reduce effort [11]. Openness and the willingness to accept that there is (always) space for improvement of the data management processes is a prerequisite. Having worked with different scientific communities, we remarked that most resources are on average FAIR (about 50% in manual assessment tools), but getting beyond

the 70% threshold involves extensive work. However, efforts towards data FAIRification are a worthwhile investment as the FAIRification is a gradual process towards improving the data quality and a FAIR data set positively affects research outcomes.

Author contributions

DW, ETI, VS and IB contributed to all sections and revised and approved the final version.

Competing interests

The authors declare that they have no competing interests.

Funding

DW and IB acknowledge funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017536 (EOSC Future) for the project on FAIR COMBINE archives. DW acknowledges funding from the German Ministry of Health (BMBF), FKZ 01KX2121.

Acknowledgement

We thank all our collaborators from our initiatives on FAIR-related infrastructures and communities including the Disease Maps, COMBINE, EOSC Future/ RDA communities.

References

1. M.D. Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* 3 (2016) 160018. doi: <https://doi.org/10.1038/sdata.2016.18>
2. RDA COVID-19 Working Group, RDA COVID 19 Case Statement. Research Data Alliance. 2020. URL: <https://www.rd-alliance.org/group/rda-covid19-rda-covid19-omics-rda-covid19-epidemiology-rda-covid19-clinical-rda-covid19-social> [accessed 2023-04-27]
3. FAIR Data Maturity Model Working Group, FAIR Data Maturity Model. Specification and Guidelines, Zenodo (2020), doi: <https://doi.org/10.15497/rda00050>
4. Waltemath et al., The first 10 years of the international coordination network for standards in systems and synthetic biology (COMBINE), *Journal of Integrative Bioinformatics* (2020) doi: <https://doi.org/10.1515/jib-2020-0005>
5. Neal et al., Harmonizing semantic annotations for computational models in biology, *Briefings in Bioinformatics* 20 (2019), doi: <https://doi.org/10.1093/bib/bby087>
6. Ramachandran et al., FAIR Sharing of Reproducible Models of Epidemic and Pandemic Forecast, *Preprints* (2022). doi: <https://doi.org/10.20944/preprints202206.0137.v1>
7. Mazein et al., Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms, *NPJ Systems Biology and Applications* (2018), doi: <https://doi.org/10.1038/s41540-018-0059-y>
8. Gawron et al., MINERVA—a platform for visualization and curation of molecular interaction networks, *NPJ Systems Biology and Applications* (2016), doi: <https://doi.org/10.1038/npjbsa.2016.20>

9. Völzke et al., Cohort Profile Update: The Study of Health in Pomerania (SHIP), International Journal of Epidemiology 6 (2022). Doi: <https://doi.org/10.1093/ije/dyac034>
10. Michaelis et al., How FAIR is NUM? - Lessons learnt from a FAIR survey within the German Network University Medicine (NUM), Proceedings of the 2023 SWAT conference, Basel, Feb. 2023.
11. Bruno et al., FAIR and Open Data in Science: The Opportunity for IUPAC, Chemistry International (2021), doi: <https://doi.org/10.1515/ci-2021-0304>

RDM Compas: Building Competencies for the Professional Curation of Research Data

Kathrin Behrens¹[\[https://orcid.org/0000-0003-1191-5110\]](https://orcid.org/0000-0003-1191-5110) and Katarina Blask²[\[https://orcid.org/0000-0003-2062-4059\]](https://orcid.org/0000-0003-2062-4059)

¹ GESIS – Leibniz Institute for the Social Sciences, Germany

² ZPID – Leibniz Institute for Psychology, Germany

Extended Abstract.

New areas of scientific research, such as research on big data, computational social sciences, or digitization and the related social changes underline the growing prevalence of data-based-research.

As the relevance of research data increases, more and more standards, processes and criteria in research data management are becoming established, not least with the justification of the FAIR [1] principles: research data should be findable, accessible, interoperable, and reusable. At the same time, the expectations for appropriate handling of research data and comprehensive research data management are increasing: for instance, the submission of a comprehensive data management plan regarding the handling of research data in research projects is now often a requirement by third-party funders or institution-specific RDM policies.

There is no doubt that a research process that is oriented (from the very beginning) to the criteria of comprehensive research data management is a key element in fulfilling the FAIR principles. However, it is equally important to recognize processes in RDM that go beyond the research process itself, such as data preparation, data archiving, and making research data available for subsequent use. These processes are usually essential components of professional data curation, which is crucial as part of an intensive research data infrastructure.

In the German social science research infrastructure, so-called research data centres (RDC) have been established for the data curation of the respective research data. There are now already 42 accredited RDCs, which are essential for archiving and providing social science research data. To meet the associated RDM requirements, curation-specific RDM competencies are mandatory.

If one focuses on the core RDM competencies of data curation, they are quite different from the RDM competencies of appropriately managing research processes. In this context, it is striking that training and continuing education for researchers in RDM topics has already been greatly expanded in recent years. In contrast, training in curation-specific RDM competencies is not yet well advanced and represents a critical gap in the development of data literacy. Accordingly, existing and future RDC staff (as well as data curators in other institutions tasked with curating social science research data) must be equipped with the appropriate RDM competencies.

We are addressing this gap in our project “Developing and exchanging RDM skills” as a working group of KonsortSWD, the Consortium for the Social, Behavioural, Educational and

Economic Sciences in the National Research Data Infrastructure (NFDI). The goal of our project is to provide comprehensive RDM competency training for existing and future staff tasked with curating (social science) research data.

Our project plan envisions several key components of competence training: the core element builds the online platform "RDM Compas: Research Data Management Competence Base", consisting of an education and training centre with modular online trainings on the one hand, and a comprehensive knowledge base covering all topics of curation-specific RDM on the other hand. In addition, a certification option is envisioned for curation-specific RDM competencies.

In the context of our presentation, due to the advanced state of work we want to focus on the knowledge base and present its structure and elements. The structural basis is a slightly simplified version of the Data Curation Lifecycle Model [2] offered by the UK Digital Curation Centre [3]. This lifecycle describes both the basic activities and the sequential process steps of data curation and therefore provides a helpful schema for teaching the necessary RDM competencies.

In order to do justice to the different subject orientations of the RDC, the contents of the knowledge base are divided into different parts: a generic part covers general RDM basics as well as the RDM contents of the sequential process steps, and a subject-specific part addresses the disciplinary specifics of the social, educational, behavioural, and economic sciences.

This structure has already been evaluated once (as of April 2023) in an extensive user study. The feedback from the testers has been taken up and implemented in a revision of the content. The usage tests have shown twofold: first, "RDM Compas" seems to be highly relevant for the target group of RDC employees and other data curators. Second, the structure of the data curation lifecycle we use is helpful in terms of orientation and content organization.

Consequently, with the knowledge base of "RDM Compas" we offer a platform that serves to close an essential gap in the RDM competence transfer: due to its explicit orientation towards the data curation lifecycle and its focus on the target group of data curators and RDC staff, it differs significantly from the previously established offerings in RDM trainings. Thus, our project makes an important contribution to the dissemination and professionalization of RDM competencies in a needs-oriented research data infrastructure.

Keywords: RDM Competencies, Data Curation, Knowledge Base

Competing interests

The authors declare that they have no competing interests.

Funding

KonsortSWD is funded by German Research Foundation as part of National Research Data Infrastructure (NFDI) under project number 442494171.

References

1. M.D. Wilkinson et.al., "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data*, vol.3, no.1, pp. 1-9, 2016, doi: <https://doi.org/10.1038/sdata.2016.18>
2. S. Higgins, "The DCC Curation Lifecycle Model". *The International Journal of Digital Curation*, vol.2, no.2, pp. 134–140, 2008, doi: <https://doi.org/10.2218/ijdc.v3i1.48>

3. "The Digital Curation Centre", <https://www.dcc.ac.uk/> (2023-04-19)