Digitalizing the Chemical Landscape:

A Comprehensive Overview and Progress Report of NFDI4Chem

Oliver Koepler [1][https://orcid.org/0000-0003-3385-4232], Christoph Steinbeck [2][https://orcid.org/0000-0001-6966-0814], Felix Bach [3][https://orcid.org/0000-0002-5035-7978], Sonja Herres-Pawlis [4][https://orcid.org/0000-0002-4354-4353], Nicole Jung [5][https://orcid.org/0000-0001-9513-2468], Johannes Liermann [6][https://orcid.org/0000-0003-2060-842X], Steffen Neumann [7][https://orcid.org/0000-0002-7899-7192], Matthias Razum [3][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].
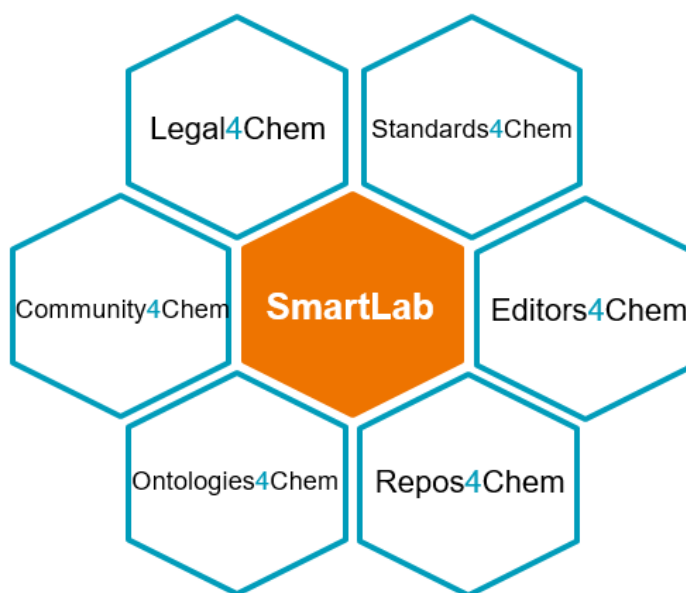
Figure 1. NFDI4Chem's Activity Clusters

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).
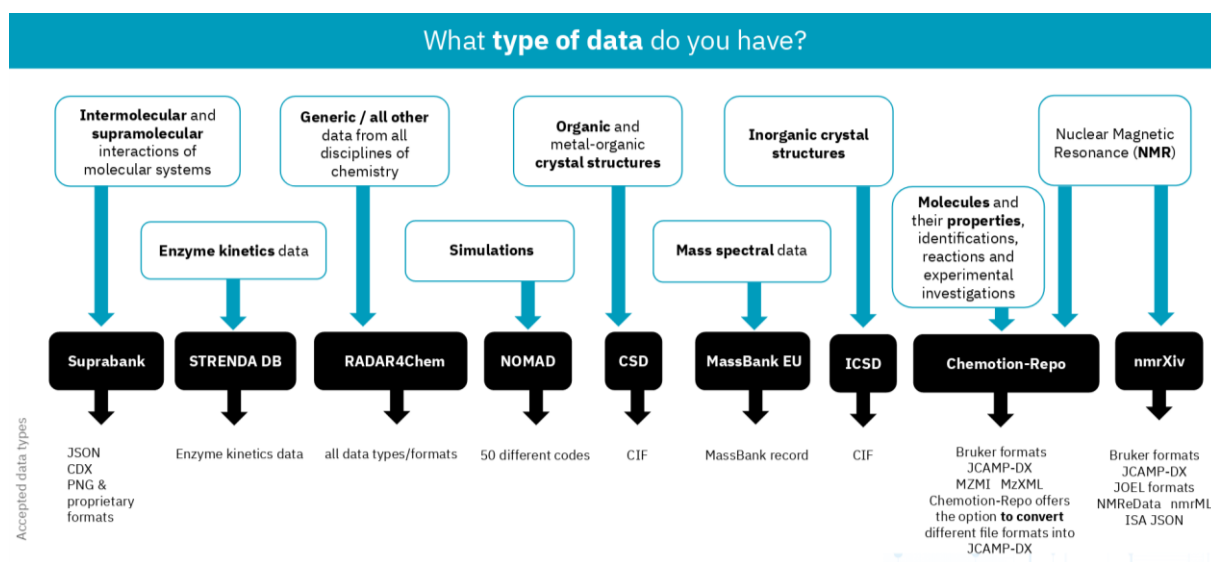

Figure 2. Data Types and Data Repositories

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

Acknowledgement

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

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

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.

6. NFDI4Chem Search Service, https://search.nfdi4chem.de, last accessed 28 April 2023