

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 'nfdi4 health Leipzig Health Atlas' logo, 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, the 'Motivation' section explains the study's goal: to analyze the impact of multimorbidity patterns and polypharmacy on the 6-month mortality rate and cognitive impairment among elderly individuals. A list of five research organizations (UNIGE, SAS, UCSC, UP, IACS) is provided, along with their specific contributions. The 'Inclusion Criteria' section lists two criteria: patients aged 65 or older and those with at least two chronic diseases. A sidebar on the right shows the 'Creators and Submitter' as Matthias Löbe and the 'Activity' section with 818 views. The bottom of the page features the 'Health Atlas - Local Data Hub/Leipzig' logo and contact information for the University of Manchester and the University of Leipzig.

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>