1st Conference on Research Data Infrastructure

Life Sciences

https://doi.org/10.52825/CoRDI.v1i.392

Published: 07 Sept. 2023

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 Golebiewski[3][https://orcid.org/0000-0002-8683-7084], Johannes Darms[1][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. questionaries 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. Undoubtfully, 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 subdomains. 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.

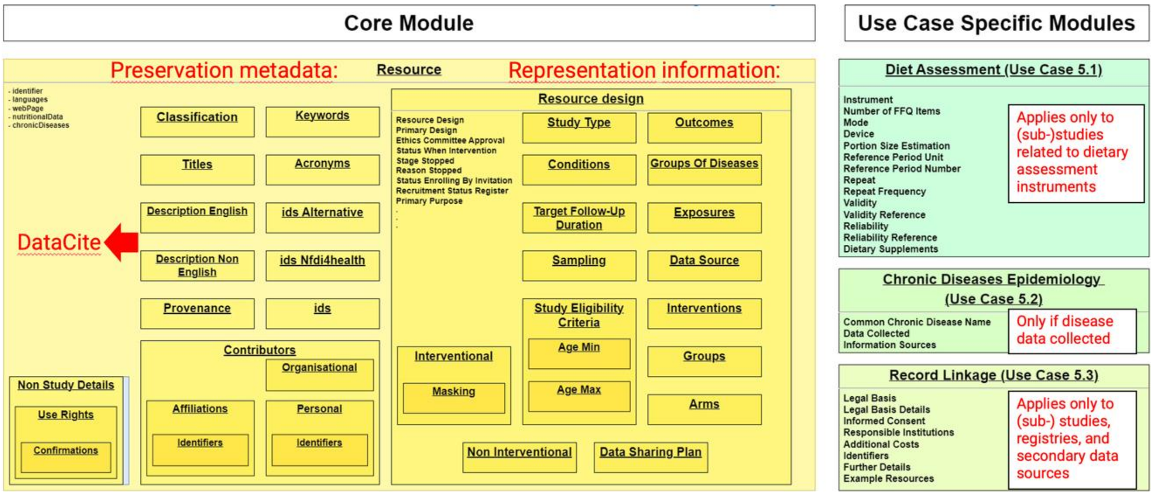
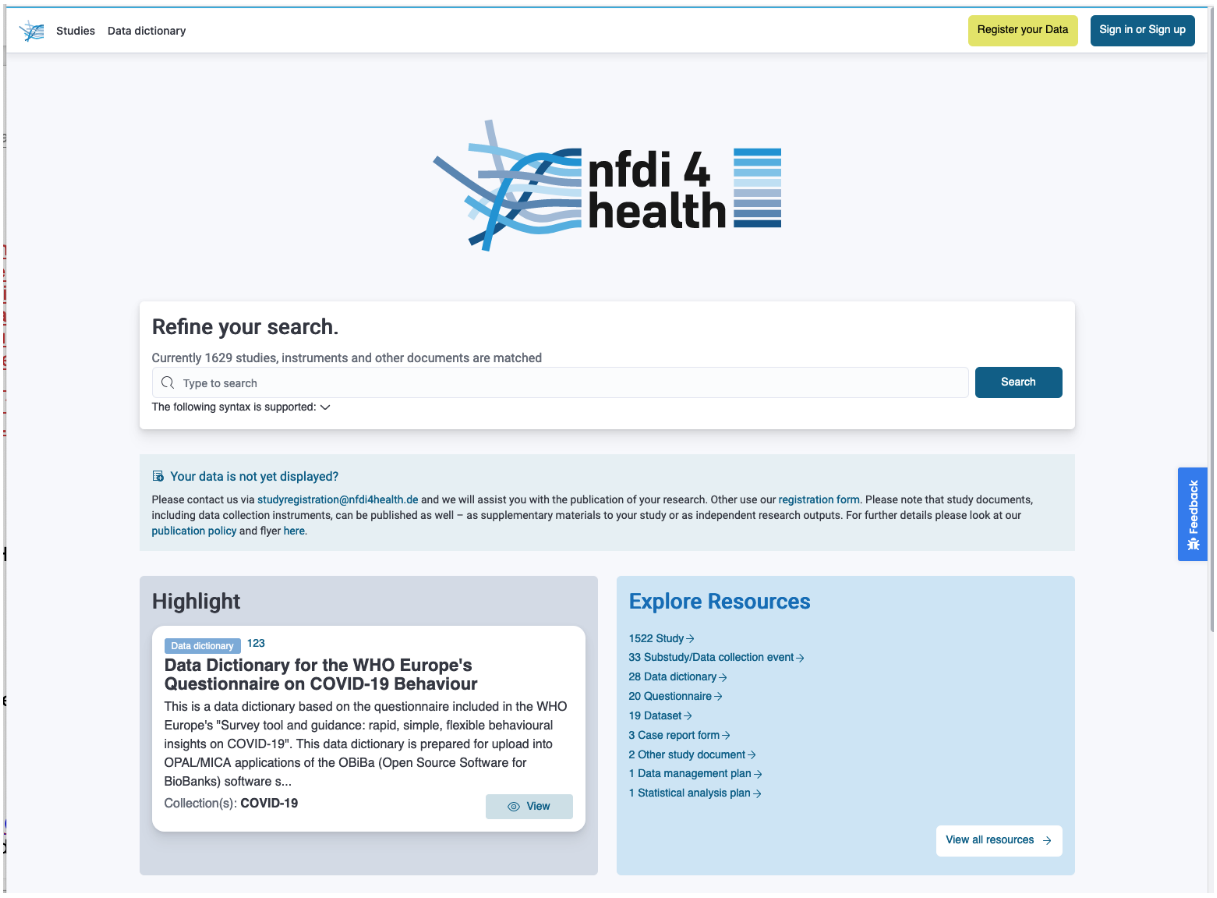


Figure 1. Subdomain-overarching NFDI4Health metadata schema designed in a modular way (schematic sketch)



Refine your search.

Currently 1629 studies, instruments and other documents are matched

Type to search

Search

The following syntax is supported:

Your data is not yet displayed?

Please contact us via studyregistration@nfdi4health.de and we will assist you with the publication of your research. Other use our registration form. Please note that study documents, including data collection instruments, can be published as well – as supplementary materials to your study or as independent research outputs. For further details please look at our publication policy and flyer here.

Highlight

Data dictionary 123

Data Dictionary for the WHO Europe's Questionnaire on COVID-19 Behaviour

This is a data dictionary based on the questionnaire included in the WHO Europe's "Survey tool and guidance: rapid, simple, flexible behavioural insights on COVID-19". This data dictionary is prepared for upload into OPAL/MICA applications of the OBiBa (Open Source Software for BioBanks) software s...

Collection(s): **COVID-19**

View

Explore Resources

1522 Study →
33 Substudy/Data collection event →
28 Data dictionary →
20 Questionnaire →
19 Dataset →
3 Case report form →
2 Other study document →
1 Data management plan →
1 Statistical analysis plan →

View all resources →

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.