J. Delgado et al. (Eds.)

doi:10.3233/SHTI210818

Improving the FAIRness of Health Studies in Germany: The German Central Health Study Hub COVID-19

Johannes DARMS[a,1], Jörg HENKE[b], Xioaming HU[c], Carsten Oliver SCHMIDT[b], Martin GOLEBIEWSKI[c] and Juliane FLUCK[a] on behalf of the NFDI4Health Task Force COVID-19
[a] *ZB MED – Information center for Live Sciences, Germany*
[b] *University Medicine of Greifswald, Germany*
[c] *Heidelberg Institute for Theoretical Studies, Germany*

Abstract. The German Central Health Study Hub COVID-19 is an online service that offers bundled access to COVID-19 related studies conducted in Germany. It combines metadata and other information of epidemiologic, public health and clinical studies into a single data repository for FAIR data access. In addition to study characteristics the system also allows easy access to study documents, as well as instruments for data collection. Study metadata and survey instruments are decomposed into individual data items and semantically enriched to ease the findability. Data from existing clinical trial registries (DRKS, clinicaltrails.gov and WHO ICTRP) are merged with epidemiological and public health studies manually collected and entered. More than 850 studies are listed as of September 2021.

Keywords. COVID-19, FAIR, study data portal

1. Introduction

A quickly growing number of clinical trials, as well as public health and epidemiological studies on COVID-19 have started and are already ongoing, but there is a lack of coordination among these efforts for securing common standards, comparable results, and – most importantly – unified access to these results.

Registries such as the German Clinical Trials Register (DRKS) and clinicaltrail.gov collect and provide information about planned, running, and completed studies. Thereby registries help researchers to find studies they are interested in. However, some studies are registered in multiple portals, while others are not registered at all. That is especially true for observational studies without any formal obligation to be registered. To our knowledge, there is no portal that provides an overarching search for clinical trials as well as for epidemiological and public health studies. In addition, current registries are commonly limited to describe overall study characteristics; detailed information about the collected data elements is lacking as well as additional study documents. This may

[1] Corresponding Author, Johannes Darms, ZB MED – Information Centre for Life Sciences, Gleueler Straße 60, 50931 Cologne, Germany, Germany; E-mail: darms@zbmed.de.

impair the possibility to find the desired information of interest. Services are needed to further facilitate the findability of studies.

The NFDI4Health Task Force COVID-19 initiative [1] was established to address those issues and increase the FAIRness – Findability, Accessibility, Interoperability and Reusability – of clinical, epidemiologic and public health studies with a COVID-19 focus. Therefore, the consortium has developed the German Central Health Study Hub Covid-19, a webservice to search for COVID-19 related studies in Germany. This service provides an overview of existing clinical trial registries (DRKS, clinicaltrails.gov and WHO ICTRP) and also includes epidemiological and public health studies that have been manually collected and entered.

2. Methods

We have combined existing platforms from different domains within the German Central Health Study Hub Covid-19. The SEEK platform [2], developed by the FAIRDOM initiative, is mainly used to store study-level metadata, as well as documents and other resources of the studies with their metadata and makes them accessible. In addition, the SEEK system is used to register data and documents with Digital Object Identifiers (DOIs). In contrast, the software systems OPAL/MICA [3] provide search and comparison techniques for data items mainly of survey instruments. OPAL provides access to characteristics of study instruments (such as labels, value lists, missing definitions and annotations). MICA allows browsing variable definitions and related studies. The variable search is enriched by semantic annotations of the Maelstrom Taxonomy [4].

To provide a unified user interface with a dedicated look & feel a new webpage has been developed. The system is a single-page application developed as a React application that combines selected information stored in SEEK, MICA/OPAL into a simplified search interface. To increase the findability and accuracy of search queries, semantically enriched information is stored in a dedicated search instance (elasticsearch).

A data model was developed to capture and harmonize information from the different sources and improve findability. The model is based on attributes used by clinicaltrails.gov [5], DRKS [6], and WHO ICTRP [7]. In addition, the required properties for assigning DOIs have been added by adhering to the DataCite scheme [8]. Those properties are needed to capture metadata about associated documents such as questionnaires, data dictionaries, and eCRFs.

A major difference of the developed data model from many others is the ability to describe a hierarchy between studies and associated documents. For example, one can model that a survey instrument is used by multiple studies (indicating reuse of data collection forms) or that one study is part of another study. A more detailed description of the data model including the set of minimal required properties and software components as well their interconnection can be found in [9]. The minimum dataset that must be included within the platform is influenced by the DataCite Schema, as some properties are required for DOIs to be assigned. In addition, some properties (title, description and study status and primary design) are mandatory for studies. The entry of other relevant metadata is recommended but not mandatory to keep the entry barrier for authors low.

The software system also includes a procedure for de-duplication of information. When duplicate resources are detected, the version from the data source with the highest priority is selected. The order is chosen by similarity with our data model i.e., more fields are equivalent. Duplicates can occur since some studies are registered in multiple registers. Especially within the WHO ICTRP dataset as it aggregates studies form other registers. The following priority list is used to resolve the problem: manually collected information, clinicaltrails.gov, DRKS, WHO ICTRP.

While not part of the software, the process to integrate study descriptions and associated documents is equality important. A business process to collect and integrate information has been designed (publication in preparation). The process is largely performed by trained data stewards. The process starts with a search for public information about a study. If some information can be obtained, it is collected and a template pre-filled with this information is sent to the study authors, otherwise a blank template is sent. Supportive mail/phone exchanges are used to assist the study author in providing needed information. When the related study documents are made available, assistance is provided in selecting the correct license.

3. Results

Our COVID-19 study hub improves the findability and accessibility of clinical, epidemiologic and public health studies related to this topic and, thus enhances their FAIRness as some of the 25 manually collected observational studies were previously not listed in any portal. The initial focus is on studies in Germany and international studies with German contributions. However, this infrastructure can also be helpful for bundling information, metadata and resources of studies in other countries or internationally, as the underlying metadata structures are generic.

Content was obtained in two ways: either by reusing existing information or by querying information directly from studies of interest, that have been identified based on a predefined requirements catalogue. Integration from existing registers (DRKS, clinicaltrails.gov and WHO ICTRP) is done automatically, but a conversion between data formats is needed. Some values may not be transferred, and others may not be provided. The final step is automatic deduplication by removing studies that are listed in multiple registers. On the other hand, the manual process of asking studies for information is labor intensive but may result in more comprehensive information in alignment with our requirements on attributes related to the study.

The study documents and resources stored in the SEEK component include study-protocol templates and data dictionaries as well as information on study-metadata structures – such as data models that describe study subjects and their clinical parameters – in addition to treatment outcomes and similar information. Additionally, direct links to primary resources and websites for the studies are included. These study information, resources and metadata can be directly searched, browsed and accessed. The MICA component of the system helps the users to find specific variables within survey instruments of interest. MICA allows to select and filter variables from the available studies to compare variable definition and its attributes.

As of September 2021, the system contained information from over 850 COVID-19 studies (46 manually collected, 158 obtained through WHO ICTRP, 468 through DRKS and 202 through NCIT) most of which are conducted in Germany. Some of the staff responsible for studies shared documents of relevance such as data collection tools, i.e., data dictionaries, questionnaires, and eCRFs. 23 data collection instruments are described at the level of individual data elements (i.e., questions, data properties), including a semantic annotation to better compare covered areas within and across instruments. The system does not contain privacy sensitive information.

The German Central Health Study Hub COVID-19 is freely accessible under https://covid19.studyhub.nfdi4health.de. The platform has already been accessed by more than 200 unique visitors a month and receives around 500 requests per day. All content can be accessed via web-interfaces and some parts are also accessible via web services (API). The software system, based on the 3 interlinked components SEEK, MICA and frontend search interface, enables browsing, accessing and comparison of COVID-19 studies and their descriptive metadata, their data collection elements, as provides search functions for studies, data collection instruments and elements, as well as related documents.

4. Discussion

We released a service to increase the FAIRness of COVID-19 related studies in Germany. The service reuses and combines existing technologies and widely used data management platforms with a sophisticated metadata schema. Data are collected and entered manually from studies (especially for epidemiological and public health studies), as well as automatically captured and reused where possible e.g., for data from clinical trials. Many interventional as well as non-interventional studies have already been published in registries such as clinicaltrails.gov and DRKS. Many studies listed in our system are taken from there. We considered the aggregation of this information as a benefit of our service. The integration of data collection instruments and item banks adds to the functionality. Item deconstruction of survey instruments and their semantic enrichment is also available in the MDM portal [10]. However, to our knowledge, there is no service that provides unified access to studies and their decomposed survey instruments.

Semantic enrichment was performed with the Maelstrom Taxonomy to ease the search for relevant information. Recent works conducted in the consortia [11] showed that SNOMED is also suited as a basis to semantically enrich data collection instruments. Therefore, we currently elaborate to do so to further increase the reusability and interoperability of the collected data.

The data schema was developed to meet the various requirements. As we could not directly use an existing schema and had to create a new one. One rationale to proceed as we did is our emphasis on fast development, due the rapid spread of the disease. However, we ensured compatibility with the current FHIR specification [12]. Our next task is to create FHIR profiles/extensions to convert our schema into a standardized format to increase the interoperability of the contained data.

The reuse of documents such as survey instruments is often hindered by legal restrictions. This problem occurs when collecting and publishing data collection instruments, there copyrights can and are claimed. Therefore, prior to inclusion in our FAIR platform, copyrights must be clarified, and documents must be licensed under some appropriate open license, such as a Creative Commons license. However, this process is not straightforward and delays the integration of instruments into the portal.

5. Conclusion

We have established a service to increase the FAIRness of clinical, epidemiologic and public health studies and associated documents. The harmonization of existing information and integration of previous unavailable information, as well as semantic enrichment of information eases the findability of COVID-19 related studies conducted in Germany. In order to further increase usefulness of the service i.e., the number of studies included, procedures to simplify the process of (meta) data collection are in preparation. The first is an interactive web-based user-form to will facilitate study registration. Furthermore, the business process used by data stewards to collect information is currently being streamlined and will be supported by the software stack. Additionally, usability is being evaluated to guide further development of the software system to meet user needs.

Acknowledgements

This work was done as part of the NFDI4Health Task Force COVID-19 (www.nfdi4health.de/task-force-covid-19-2). We gratefully acknowledge the financial support of the German Research Foundation (DFG) – Project Number 451265285, and from the Klaus Tschira Foundation (KTS).

References

[1] Task Force COVID-19 Team. Task Force COVID-19 - NFDI4Health [Internet]. [cited 2021 Aug 6]. Available from: https://www.nfdi4health.de/de/task-force-covid-19

[2] Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, et al. SEEK: a systems biology data and model management platform. BMC systems biology. 2015;9(1):1–12.

[3] 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. 2017;

[4] Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I. Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. PLoS One. 2018;13(7):e0200926.

[5] National Library of Medicine, National Institutes of Health. XML Schema for ClinicalTrials.gov public XML [Internet]. [cited 2021 Aug 6]. Available from: https://clinicaltrials.gov/ct2/html/images/info/public.xsd

[6] DRKS. Description of entry fields [Internet]. [cited 2021 Aug 6]. Available from: https://www.drks.de/drks_web/navigate.do?navigationId=entryfields&messageDE=Beschreibung%20der%20Eingabefelder&messageEN=Description%20of%20entry%20fields

[7] WHO. World Health Organisation - ICTRP Search Portal [Internet]. [cited 2021 Aug 6]. Available from: https://www.who.int/clinical-trials-registry-platform/the-ictrp-search-portal

[8] DataCite Metadata Working Group. DataCite metadata schema documentation for the publication and citation of research data. Version 4.3 [Internet]. [cited 2021 Aug 6]. Available from: https://schema.datacite.org/meta/kernel-4.3/

[9] Schmidt CO, Darms J, Shutsko A, Löbe M, Nagrani R, Seifert B, et al. Facilitating Study and Item Level Browsing for Clinical and Epidemiological COVID-19 Studies. Studies in Health Technology and Informatics. 2021;281:794–8.

[10] Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database. 2016;2016.

[11] Vorisek CN, et al. Evaluating Suitability of SNOMED CT in Structured Searches for COVID-19 Studies. In: Public Health and Informatics. IOS Press; 2021. p. 88–92.

[12] Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE international symposium on computer-based medical systems. IEEE; 2013. p. 326–31.