Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210284

Facilitating Study and Item Level Browsing for Clinical and Epidemiological COVID-19 Studies

Carsten Oliver SCHMIDT[a,1], Johannes DARMS[b], Aliaksandra SHUTSKO[b], Matthias LÖBE[c], Rajini NAGRANI[d], Bastian SEIFERT[d], Birte LINDSTÄDT[b], Martin GOLEBIEWSKI[e], Sofiya KOLEVA[f], Theresa BENDER[g], Christian Robert BAUER[g], Ulrich SAX[g], Xiaoming HU[e], Michael LIESER[e], Vivien JUNKER[e], Sophie KLOPFENSTEIN[h], Atinkut ZELEKE[a], Dagmar WALTEMATH[a], Iris PIGEOT[d] and Juliane FLUCK[b], on behalf of the NFDI4Health Task Force COVID-19

[a]University *Medicine of Greifswald*
[b]*ZB MED – Information Centre for Life Sciences*
[c]*University of Leipzig*
[d]*Leibniz Institute for Prevention Research and Epidemiology - BIPS*
[e]*Heidelberg Institute for Theoretical Studies*
[f]*McGill University Health Centre*
[g]University *Medical Center Göttingen*
[h]*Berlin Institute of Health*

Abstract. COVID-19 poses a major challenge to individuals and societies around the world. Yet, it is difficult to obtain a good overview of studies across different medical fields of research such as clinical trials, epidemiology, and public health. Here, we describe a consensus metadata model to facilitate structured searches of COVID-19 studies and resources along with its implementation in three linked complementary web-based platforms. A relational database serves as central study metadata hub that secures compatibilities with common trials registries (e.g. ICTRP and standards like HL7 FHIR, CDISC ODM, and DataCite). The Central Search Hub was developed as a single-page application, the other two components with additional frontends are based on the SEEK platform and MICA, respectively. These platforms have different features concerning cohort browsing, item browsing, and access to documents and other study resources to meet divergent user needs. By this we want to promote transparent and harmonized COVID-19 research.

Keywords. Browsing metadata, COVID-19, FAIR data, metadata standards

1. Introduction

COVID-19 poses a major challenge to individuals and societies around the world. Lockdown efforts dramatically change social lives, economic prospects and health services. This creates a huge demand for scientific data to understand the virus' spread,

[1] Corresponding author, University Medicine Greifswald, Institute for Community Medicine, SHIP-KEF, Walther-Rathenau Str. 48. 17475 Greifswald; E-mail: Carsten.schmidt@uni-greifswald.de

therapeutic options and consequences of the pandemic. Knowing about ongoing research activities is indispensable to better align novel with existing research activities and to avoid a waste of resources. Though systematic overviews are readily available for trials metadata using registries such as the International Clinical Trials Registry Platform (ICTRP) [1], the situation is much more complicated for epidemiological and public health studies. While some registries are available, such as a registry on seroepidemiological studies [2] or the COVID-19 research registry [3], the scope and depth is mostly limited. Improvements are necessary to bridge different fields of medical research with harmonized search options. Information should be accessible beyond mere study descriptions, such as study protocols, statistical analysis plans, or instruments.

The NFDI4Health Task Force COVID-19 initiative, an interdisciplinary German network project within the National Research Data Infrastructure (NFDI) initiative in Germany, targets these shortcomings [4]. This paper describes a consensus metadata model to facilitate the integration of study-related information along with its implementation in complementary web-based platforms to meet needs of divergent users.

2. Methods

We created a health studies data model that is able to (1) integrate information from trials, epidemiological and public health studies, while (2) being largely compatible with data models in trials registries, e.g. ICTRP [1], German Clinical Trials Register (DRKS) [5], and the Minimum Information About BIobank data Sharing (MIABIS) [6]. To better accommodate content from epidemiological studies, elements from the Maelstrom data model were incorporated [7]. In addition, we integrated elements from the DataCite Metadata Schema to improve citability and searchability [8]. A mapping [9] was conducted against HL7 FHIR [10] and CDISC ODM [11] to facilitate interoperability. The interlinked platforms are:

(1) a Central Search Hub (CSH) for study level information. It consists of a user interface developed as a single-page application (SPA) in the react framework and a RESTful API service. The data itself is stored in an Elasticsearch Instance and exposed via the web-service. SPA uses a REST interface.

(2) the COVID-19 study hub SEEK platform links COVID-19 studies with their metadata, documents (assets) and other information. The underlying SEEK [12, 13] is a well-established life science data and model management platform developed by the FAIRDOM initiative and built as open-source software using Ruby on Rails. It can be accessed manually via a browsing interface or via a JSON-based API web-service.

(3) Opal and MICA, open-software solutions for epidemiological data management add options to browse items from survey instruments and item banks [14]. Opal and MICA are interoperable web-applications, written in Java, JavaScript and PHP. Opal is used to store information on study variables with semantic annotations using the Maelstrom taxonomy [7]. MICA is a metadata catalogue and data discovery tool.

3. Results

Figure 1 provides an overview of the health studies data model, more detailed information is in the metadata schema [15]. The model is generic in its ability to represent a wide range of resources such as studies, sub-studies, datasets, as well as diverse study

resources and documents, among others. Its application is not restricted to COVID-19 studies but pursues generic applicability. The hierarchical relation between resources allows for an organization of studies with complex designs such as multiple data collection events. The compatibility with DataCite enables the assignment of a Digital Object Identifier (DOI) guaranteeing persistent access to and the findability of published resources. A selected license allows the legitimate (re-)use of contributions. The database contains trial metadata from ICRTP, DRKS, and manually collected data on observational studies and documents with 581 studies as of March 3rd.

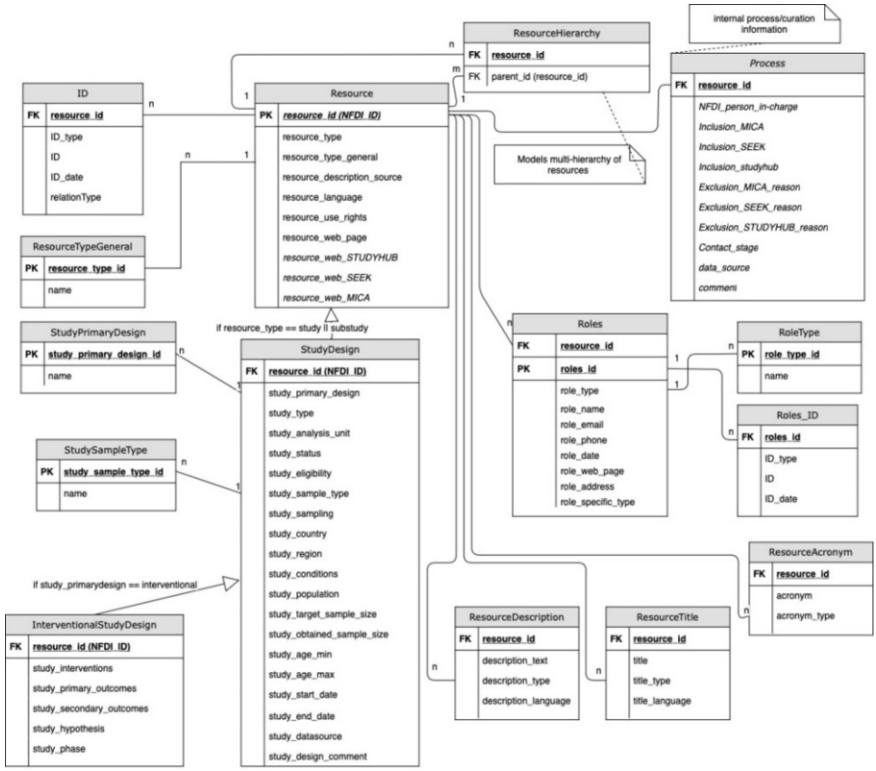

Figure 1. Simplified entity relationship diagram of the health studies data model.

The CSH [16] enables to search for studies and for related study documents (Figure 2) based on information represented in the health studies data model. In addition to overviews provided by structured metadata templates and the hierarchical structure of associated resources, additional connections to specialized sub-components are given. The CSH acts as a broker between them: the system MICA allows for browsing and exploring study instruments in detail, while SEEK provides access to and versioning of interrelated study assets (documents and resources).

The configurable architecture and data model of the COVID-19 study hub SEEK [17] allows for complex and relational sharing of a wide range of resources and objects such as studies, their protocols and documents, as well as their (meta-)data, models, and results. The metadata model was implemented in SEEK via adapting existing structuring and classification mechanisms (ISA structure, [12]) and using a novel feature for creating customs metadata. SEEK is used as both, as metadata catalogue and centralized

repository. Study assets are either uploaded or only registered and stored externally, but are accessible in a coherent way, independently of their internal or external storage. The system allows for storage of study assets in different stages and thus versioning. Studies and assets have persistent URLs enabling references. It is also possible to generate a DOI for any asset that is public and visible.

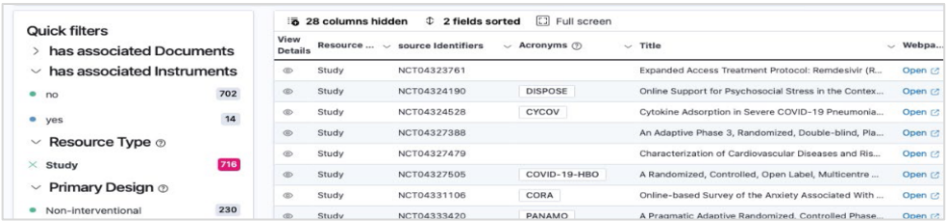

Figure 2. Central Search Hub sample view.

The major focus of MICA [18] is on items from survey instruments, which are nested within studies. A searchable semantic annotation of survey items within 18 domains and 135 sub-domains facilitates comparisons of items across studies (Figure 4). While CSH and SEEK pursue a broad display of COVID-related studies, the focus of MICA is on selected resources with their instruments, e.g. to guide the creation of new surveys.

Table 1. Differential focus of the three web-portals in the COVID-19 hub

Feature	CSH	SEEK	MICA
Simple study level relational metadata overview	*		
Study and document search and grouping	*		
Information brokering across the three portals	*		
Complex relational structuring of studies, documents, resources		*	
Assignment of DOIs		*	
Versioning of study resources and documents		*	
Browsing of item* level metadata across resources			*
Searchable semantically annotated items*			*
Download of item* selections across resources			*

* Item here refers to data elements, e.g. study variables such as a question in a survey.

4. Conclusion

Due to imminent danger posed by the virus, traditional quality assurance mechanisms had to be shelved in favor of rapid study starts, increasing the risk of doubling and ignoring activities. We introduce an infrastructure to help researchers gain an overview of medical research activities related to COVID-19. Data from existing trials registries such as the ICTRP is combined with information from epidemiological and public health studies. Other resources can be handled as well. We did not follow a one size fits all users approach but rather exploited strengths of different views on studies and documents to meet divergent user demands while retaining the advantage of a single database backend. We present work in progress and the content is expanded continuously. Challenges are the manual addition and curation of novel metadata. Based on user feedback, further tasks are to optimize the interplay between the platforms. Both, MICA and SEEK are customizable so that further semantics like SNOMED CT can be used in future versions. This and more dynamic links to the metadata model would benefit the future presentation of content. While the current focus is on German studies and COVID-19, all developments are generic in being usable beyond this application scenario.

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, PI 345/17-1/SCHM 2744/9-1, and from the Klaus Tschira Foundation.

References

[1] World Health Organisation. ICTRP Search Portal [Internet]. Geneva: WHO; 2021, Available at https://www.who.int/clinical-trials-registry-platform/the-ictrp-search-portal, Accessed Jan 13, 2021.
[2] Robert Koch Institute. Seroepidemiological studies in Germany [Internet]. Berlin: RKI; 2021, Available at https://www.rki.de/EN/Content/infections/epidemiology/outbreaks/COVID-19/AK-Studien-english/Sero_List.html, Accessed Jan 13, 2021.
[3] American Society for Microbiology. COVID-19 Research Registry [Internet]. Washington, DC: The American Society for Microbiology; 2021, Available at https://asm.org/COVID/COVID-19-Research-Registry/Epidemiology, Accessed Jan 13, 2021.
[4] NFDI4Health. Task Force COVID-19 [Internet]. [place unknown]: NFDI4Health; 2020, Available at https://www.nfdi4health.de/task-force-covid-19-2/, Accessed Jan 13, 2021.
[5] DRKS. Description of entry fields [Internet]. Cologne: German Clinical Trials Register; 2019, Available at https://www.drks.de/drks_web/navigate.do?navigationId=entryfields&messageDE=Beschreibung%20der%20Eingabefelder&messageEN=Description%20of%20entry%20fields, Accessed Jan 13, 2021
[6] Merino-Martinez R, Norlin L, van Enckevort D, Anton G, Schuffenhauer S, Silander K, Mook L, et al. Toward global biobank integration by implementation of the Minimum Information About BIobank data Sharing (MIABIS 2.0 Core). Biopreserv Biobank. 2016;14:298-306.
[7] Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I. Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. PLoS One. 2018 Jul 24;13(7):e0200926.
[8] DataCite Metadata Working Group. DataCite metadata schema documentation for the publication and citation of research data. Version 4.3 [Internet]. [place unknown]: DataCite e.V.; 2019 Aug, Available at: https://schema.datacite.org/meta/kernel-4.3/, Accessed Jan 13, 2021.
[9] NFDI4Health Task Force COVID-19 Metadata Schema Mapping [Internet]: FAIRDOMHub; 2021, Available at: https://fairdomhub.org/data_files/3973, Accessed Jan 13, 2021.
[10] HL7 FHIR. Documentation Index [Internet]. [place unknown]: HL7.org; 2019, Available at http://hl7.org/fhir/documentation.html, Accessed Jan 13, 2021.
[11] Clinical Data Interchange Standards Consortium. Data Exchange [Internet]. [place unknown]: CDISC; 2020, Available at https://www.cdisc.org/standards/data-exchange, Accessed Jan 13, 2021.
[12] Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, Weidemann A, et al. SEEK: a systems biology data and model management platform. BMC Syst Biol. 2015 Jul 11;9:33.
[13] Wolstencroft K, Krebs O, Snoep JL, Stanford NJ, Bacall F, Golebiewski M, Kuzyakiv R, et al. FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. Nucleic Acids Res. 2017 Jan 4;45(D1):D404-7.
[14] 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. Int J Epidemiol. 2017 Oct 1;46(5):1372-8.
[15] Golebiewski M, Löbe M, Schmidt CO, Lehne M, Shutsko A, Darms J. NFDI4Health Task Force COVID-19 Metadata Schema [Internet]: FAIRDOMHub; 2021, Available at: https://fairdomhub.org/data_files/3972?version=1, Accessed Jan 13, 2021.
[16] NFDI4Health Task Force COVID-19 Central Search Hub, Available at: https://covid19.studyhub.nfdi4health.de
[17] NFDI4Health Task Force COVID-19 SEEK Platform, Available at https://seek.covid19.studyhub.nfdi4health.de
[18] NFDI4Health Task Force COVID-19 MICA Platform, Available at https://mica.covid19.studyhub.nfdi4health.de