The NFDI4Health – Task Force COVID-19

Iris Pigeot[1][https://orcid.org/0000-0001-7483-0726], Juliane Fluck[2,3 [https://orcid.org/0000-0003-1379-7023]], Johannes Darms[2][https://orcid.org/0000-0001-5809-2276], and Carsten Oliver Schmidt[4][https://orcid.org/0000-0001-5266-9396] on behalf of the NFDI4Health – Task Force COVID-19

[1] Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany

[2] Information Centre for Life Sciences, Köln, Germany

[3] University of Bonn, Bonn, Germany

[4] Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

Keywords: First keyword, Second, Third keyword

COVID-19 posed one of the greatest challenges to individuals and societies worldwide in recent decades. Public health research, epidemiological and clinical studies were essential to track the spread of SARS-CoV-2 responsible for the pandemic and its variants, to better understand the consequences for health and social life, and to identify effective treatment and vaccination methods. Such studies provided policy makers, industry, health care providers, and society with an empirical basis for containing and managing the pandemic and for making decisions that were based on the most recent data. Therefore, the COVID-19 pandemic excellently illustrates the relevance of data sharing and the importance of providing an effective infrastructure. From the researchers' perspective, there were significant challenges associated with this request. In a very short time, numerous projects, studies, and networks had emerged to investigate the pandemic, making it increasingly difficult to maintain an overview. Such an overview would have been essential to coordinate research activities, avoid unplanned duplication of research, and to implement studies in a harmonized manner.

Hurdles already varied vastly when trying to find ongoing studies. Due to the existing obligation to register clinical trials in registries, their well-structured metadata are available. In contrast, the situation for epidemiological and public health studies was much less clear. Although there were several national and international overviews on the Internet, e.g., a COVID-19 research registry of the American Society for Microbiology [1], the COVID-19 research overview of the Medical Informatics Initiative [2], or the German Data Forum [3], these overviews were inconsistent in scope, timeliness, and depth of information.

It became even more difficult when a detailed insight into protocols, survey instruments, item banks, and other study documents was requested across studies. Only in isolated cases did projects provide access to relevant information. For example, the German Corona Consensus Dataset (GECCO) is a positive example of harmonized data collection based on international medical IT standards using a coordinated core dataset. It was created in the Network University Medicine (NUM) [4], which primarily coordinates hospital-related research [5].

In addition, during the course of the pandemic, other obstacles became apparent that hampered efficient research: Although in some cases the same individuals were included in different studies and further health data of these individuals were stored, e.g., by health insurance companies, there are insufficient options to link these data on an individual level. This limits the possibilities of obtaining a sufficiently comprehensive picture of disease occurrence, on progression prognosis or vaccination consequences.
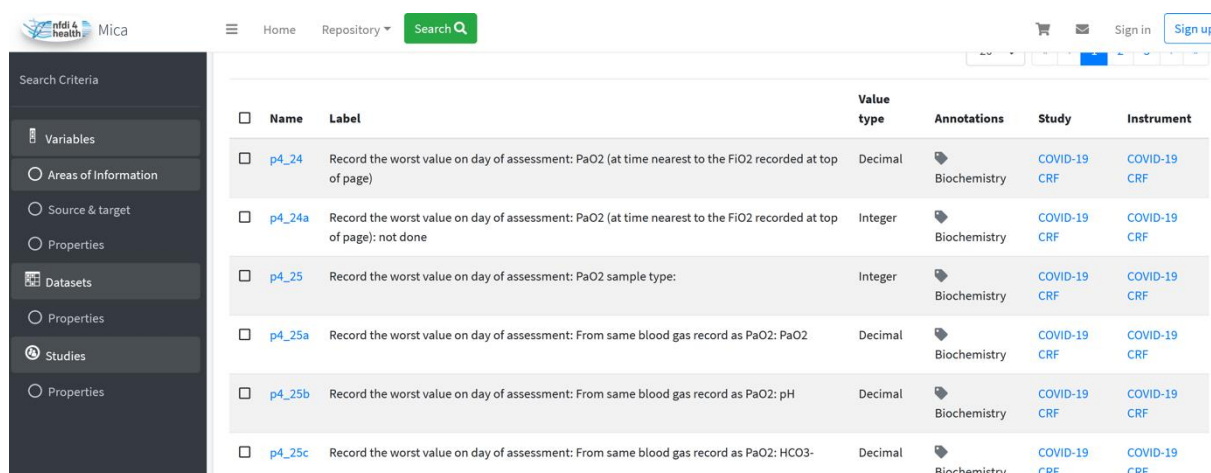
There was also a need to share even preliminary research results in order to respond adequately to the pandemic. As a result, the publication of preprints and the importance of preprint servers for sharing results without peer review increased, as the established peer-reviewed publication process was to slow during this period.

In summary, despite positive examples, the German clinical studies on COVID-19 and corresponding datasets in epidemiology and public health have only insufficiently met the requirements of the so-called FAIR principles [6]. To address this shortcoming, the NFDI4Health Task Force COVID-19 [7] was established as part of NFDI4Health [8]. In doing so, the NFDI4Health Task Force COVID-19 focused at research related to patients with COVID-19 as well as the public health consequences of the pandemic outbreak on the general population.

This project led to a range of key outcomes that meanwhile serve as a basis for NFDI4Health developments. For one part, a first version of a central search portal, the German Central Health Study Hub COVID-19, has been established. This application allows users to search, explore and retrieve study-related information. The main focus of the application is the consolidation of distributed information in order to reduce search time. It has since been expanded to include additional datasets and functionality to form the NFDI4Health German Central Health Study Hub (https://csh.nfdi4health.de) (Figure 1). The search engine is complemented by an instrument portal function. This allows COVID-19 related survey instruments to be searched semantically. The Maelstrom taxonomy [11],[12] is used for this purpose. This taxonomy consists of 18 domains, such as socio-demographic and economic characteristics and diseases, which in turn are divided into 135 subdomains (e.g. ICD domains). The following link (https://mica.covid19.studyhub.nfdi4health.de/) provides direct access to the function.

Figure 1. German Central Health Study Hub – COVID-19 collection

An important underlying activity was the development of a metadata model to describe studies and related resources. Compatibility with existing registries and standards such as the ICTRP [9] and the German Register of Clinical Studies (DRKS) [10], the Minimum Information About Biobank Data Sharing (MIABIS), the Maelstrom data model and the DataCite metadata schema was taken care for. Mappings against HL7 FHIR and CDISC ODM standards were conducted. These results have also been incorporated into the NFDI4Health initiative and are already being expanded.

Figure 2. Instrument portal

The initiative also set up and elaborated a range of other services such as tools for assessing imaging quality, tools for assessing data quality, and a semantic search engine for preprints [13]. Regarding the latter, metadata were queried from the preprint servers medRxiv, bioRxiv, ChemRxiv, ResearchSquare, arXiv and Preprints.org and transferred to a shared data schema. A terminology was created to identify viral SARS-CoV-2 proteins using a dictionary-based algorithm, a web-based user interface, and a programming interface were developed to provide users with semantic search functionalities.

This infrastructure makes it easier to find research and its results on SARS-CoV-2 and COVID-19 from public health, epidemiology and clinical studies. The described developments, which were initiated as part of the NFDI4Health Task Force COVID-19 [12] are also relevant beyond COVID-19, as the challenges addressed are generic for finding and exploiting research data. Thus, the Task Force COVID-19 may be regarded as a kind of microcosm that served as a blueprint for the activities required to successfully implement the services and standards developed by the NFDI4Health.

Author contributions

All authors collaborate in the NFDI4health project. IP, COS prepared the manuscript, all authors reviewed and finalized.

Competing interests

The authors declare that they have no competing interests.

Funding

References

1. American Society for Microbiology (2022), COVID-19 research registry. https://asm.org/COVID/COVID-19-Research-Registry/Epidemiology. Last access: April 15, 2023
2. Medical Informatics Initiative (2021), Overview of Covid-19 research. https://www.medizininformatik-initiative.de/en/node/410. Last access: April 15, 2023

3. German Data Forum (2023), Studien zur Corona-Pandemie. https://www.konsortswd.de/ratswd/themen/corona/studien/. Last access: April 15, 2023

4. Network University Medicine (2023), Homepage. https://www.netzwerk-universitaetsmedizin.de/. Last access: April 15, 2023

5. Sass, J., Bartschke, A., Lehne, M. et al. (2020). The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak 20, 341. DOI: https://doi.org/10.1186/s12911-020-01374-w

6. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. DOI: https://doi.org/10.1038/sdata.2016.18

7. NFDI4Health (2023), Task Force COVID-19. https://www.nfdi4health.de/en/task-force-covid-19.html. Last access: April 15, 2023

8. NFDI4Health (2023), NFDI4Health – Eine Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten. https://www.nfdi4health.de/en/. Last access: April 15, 2023

9. World Health Organization (2023), ICTRP search portal. https://www.who.int/clinical-trials-registry-platform/the-ictrp-search-portal. Last access: April 15, 2023

10. German Clinical Trials Register (DRKS) (2023), Deutsches Register Klinischer Studien. https://www.drks.de/. Last access: April 15, 2023

11. Schmidt, C. O., Fluck, J., Golebiewski, M., et al. (2021). COVID-19-Forschungsdaten leichter zugänglich machen – Aufbau einer bundesweiten Informationsinfrastruktur [Making COVID-19 research data more accessible-building a nationwide information infrastructure]. Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz, 64(9), 1084–1092. DOI: https://doi.org/10.1007/s00103-021-03386-x

12. Schmidt CO, Darms J, Shutsko A, et al. (2021). Facilitating Study and Item Level Browsing for Clinical and Epidemiological COVID-19 Studies. Studies in Health Technology and Informatics. 2021 May;281:794-798. DOI: 10.3233/shti210284. PMID: 34042687.

13. Langnickel, L., Darms, J., Heldt, K., et al. (2022). Continuous development of the semantic search engine preVIEW: from COVID-19 to long COVID. Database, Volume 2022, baac048. DOI: https://doi.org/10.1093/database/baac048