Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200189

Problems in FAIRifying Medical Datasets

Matthias LÖBE[a,1], Franz MATTHIES[a], Sebastian Stäubert[a], Frank A. Meineke[a], and Alfred WINTER[a]
[a] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), Leipzig University, Leipzig, Saxony, Germany*

Abstract. Despite their young age, the FAIR principles are recognised as important guidelines for research data management. Their generic design, however, leaves much room for interpretation in domain-specific application. Based on practical experience in the operation of a data repository, this article addresses problems in FAIR provisioning of medical data for research purposes in the use case of the Leipzig Health Atlas project and shows necessary future developments.

Keywords. FAIR data, Research data management, Data sharing

1. Introduction

The FAIR Guiding Principles for scientific data management have been widely accepted since their publication in 2016 [1]. Nevertheless, their practical implementation is difficult, especially in the area of particularly sensitive personal medical data. While many data repositories and data management solutions claim to be FAIR, the qualitative implementation of the FAIR principles varies [2]. A typical university hospital supports various types of research projects in addition to patient care, such as clinical trials for the approval of new drugs or therapy optimization, disease-specific registries for the most complete possible registration of all patients suffering from a specific disease, population-based cohorts for epidemiological long-term monitoring of the population, biomaterial banks, molecular genetic analysis (OMICS) and much more. Most of these research projects generate high quality data in terms of completeness, consistency and accuracy. Data is collected, curated, analyzed; but rarely used again. The objective of this contribution is to show the problems that arise when building a research data repository in accordance with the FAIR principles, with special emphasis on the health research domain.

2. Method

Medical research on humans is resource-intensive and ethically challenging, so it should be made possible to get the maximum out of the data. The Leipzig Health Atlas is a research data repository that has been built up over the last few years [3] and is now constantly being filled with content. Based on the SEEK platform [4], not only

[1] Corresponding Author: Matthias Löbe, Härtelstraße 16-18, 04107 Leipzig, Germany; E-mail: matthias.loebe@imise.uni-leipzig.de

data sets but also other types of research artefacts such as systems biology pipelines, bioinformatics models or medical ontologies will be managed. In addition to the core task of long-term archiving data are to be made available to other researchers ("Data Sharing“) and analyses of data should be repeatable by third parties ("Reproducibility“). The method empirical-analytical approach used here is based on our experience in operating the portal and the consulting and feedback of scientists when uploading their results.

3. Results

This chapter outlines various FAIR hurdles for biomedical research, broken down according to the main axes of the FAIR principles, which have arisen several times and for which proposals are made for dismantling them.

3.1. Considerations for FAIR axis 1: Findability

Findability deals with the problem of recognising the existence of resources and referencing them. Many data repository software systems or content management systems assign (locally) unique numbers by themselves, which could be made globally unique by a suitable prefix. However, the required "eternal persistence" cannot be guaranteed with regard to updates or system changes. This results in the need for special digital identifier registry software. Such systems already exist in the medical field, e.g. the HL7 Object Identifier (OID) Registry as defined in the recommendation ITU-T X.660 and ISO/IEC 9834 series. Unfortunately, the OID system [5] is only one of many possible handle systems, Digital Object Identifiers (DOI) are widely used in scientific journals, and Uniform Resource Identifiers (URI) are used in distributed web-based software systems. While each of these systems can be designed to be *resolvable* in principle, there is no agreement on other desirable characteristics [6] such as descriptive human readable designation as part of identifiers.

Another problem is the granularity at which data is to be assigned its own identifiers. It is undoubtedly important, for example, to provide a data set that was the basis of a scientific publication in one version with its own identifier. However, medical databases are characterized by frequent and extensive changes, whereby data is not only added, but also updated. A continuous versioning would be impracticable here. A further problem is that it is desirable with regard to the use of data by external systems to also be able to reference only parts of data sets or single data elements. In the sense of the idea of a Semantic Web, each information atom would get its own URI, which is, however, insufficiently supported in practice by existing software.

While the actual search can be very well supported by existing software using keywords, ontologies, or categorical filters, there is a lack of community accepted metadata vocabularies for describing data sets (such as MIAME – Minimum information about a microarray experiment or the WHO Clinical Trial Registration Data Set) for many areas of biomedical research, which often makes it difficult for users to navigate within a repository. Furthermore, due to the growing number of repositories, it would be desirable if these were federatively linked so that a user does not have to formulate a request several times in different places and aggregate the results. Here, too, there is a lack of a unifying data catalog specification for available health data describing the content of data collections and the location. Ultimately, individual data elements must be described very precisely in order to prevent misinterpretation. Simple designations cannot do this. Metadata repositories (MDR) allow semantic specification of clinical concepts through annotation with medical terminologies, as well as the precise specification of units of measurement, formats, data types, and reference ranges. Different systems exist, but their use is very complex and does not fit into the regular work processes, which limits their application [7].

3.2. Considerations for FAIR axis 2: Accessibility

Accessibility enables a researcher to actually retrieve data. To have metadata permanent available, even when the data are no longer available, is quite easy to achieve technically but harder to enforce culturally because some researchers have not yet fully embraced the meaning of this criterion. Access via a standardized communication protocol is also not a technical problem since Representational State Transfer (REST) interfaces based on the HTTP protocol have established themselves and current open standards such as HL7 Fast Healthcare Interoperability Resources (FHIR) [8] are based on them. The difficulties with access to personal medical data lie rather in their sensitive nature for the individual patient. Comprehensive regulations such as the European General Data Protection Regulation (GDPR) regulate in detail the conditions and restrictions for their collection and further processing. In general, the use of such data for research purposes requires the informed consent of the patient or proband. The declaration of consent must be sufficiently specific with regard to the research objective, the persons accessing the data and the circumstances of the data processing, which can prevent subsequent data sharing. Also, identifying data must not be stored together with other research data, which makes subsequent data collection or record linkage more difficult. Furthermore, participation in research projects is always voluntary and can be recalled at any time. Anonymisation procedures for individual medical data can remove these restrictions, but are accompanied by a great loss of information entropy. For this reason, it is necessary to create and use further harmonised metadata vocabularies on topics such as the legal basis for data collection or the different variants of Informed Consent.

Since such data will rarely be automatically accessible, the process of requesting data access and secure data retrieval plays an important role. In practice, a simple reference to a contact person for each data record represents a major hurdle for both the data holder and the interested party, since medical researchers are not always fully aware of the legal framework. The establishment of a Data Access Board for checking and approving data use proposals according to a usage concept is strongly recommended. The data repository should provide an authentication service for internal and external users based on a rights and roles concept. To address the concerns of data owners about possible breaches of data privacy during data sharing, contact points for questions regarding data sharing, data protection, etc. should be established.

Further problems regarding permanent accessibility are contradictory rules for the storage of data. While individual funders consider a storage period of at least 10 years as part of good scientific practice to be appropriate, some data usage agreements state that the shared data must be deleted immediately after the end of the project, which has already led to data records no longer being available shortly after the publication of a scientific article. A regulatory harmonisation must take place here.

3.3. Considerations for FAIR axis 3: Interoperability

Interoperability is the ability of distributed, heterogeneous systems to exchange information in an unambiguous and exploitable manner. There are countless standards, conventions and best practices in biomedical research, but in many cases researchers take advantage of the freedom of science to act according to their own ideas. For the operators of research data repositories, this often means that only minimal ideas about data structures can be enforced (e.g. a tabular format with character-separated values).

Since technical and semantic interoperability are generally subject to higher requirements, operators should support data owners in transforming the data into a widely applicable format for knowledge representation. Such a FAIRification process requires the definition of a rich target data model. Various common data models are suitable for this; the selection can be influenced by later intentions of use. The latest, but also most universally applicable model is the already mentioned HL7 FHIR. It covers practically all areas of medicine, even if some resources are not yet finally specified. In the area of clinical trials, the standards of the CDISC organization are of importance, especially the Operational Data Model (ODM) [9] in conjunction with the Study Data Tabulation Model (SDTM). For large, distributed data analyses, the OHDSI Observational Medical Outcomes Partnership (OMOP) [10] model has gained great popularity. For all these models, however, rudimentary ETL tools are available at best.

International medical terminologies such as ICD-10, LOINC or SNOMED CT are very suitable for describing clinical concepts in detail. They only partially meet the requirement that vocabularies should also comply with FAIR principles. Terminology services can aid the linking process.

3.4. Considerations for FAIR axis 4: Reusability

Reusability describes the application of data for secondary purposes independent of the original use. With regards to medical data, provenance is the most important topic of the FAIR reusability metric. While provenance is a broad topic and the demarcation to data acquisition is not sharp, data owners should specify in detail the circumstances of the data collection and processing, i.e. data sources, data validation rules, format conversions, data cleansing, derived or aggregated data, measuring instruments, scripts, software libraries, observers. This greatly increases the confidence of external researchers in the data sets. A further recommendation is the provision of simple web-based visual analysis tools such as tranSMART [11], which give potential interested parties an overview of the depth of the available data and so enhances reusability.

3.5. Additional considerations for medical FAIR data

Some aspect does not fit into one of the FAIR data categories but are nevertheless part of our recommendations: 1) Data quality is a primary concern when one relies on external data. All procedures to ensure high data quality (technical validations, manual curation) should be made explicit. 2) If datasets are not eligible for sharing (privacy, volume), privacy-preserving data analysis techniques could be an option. 3) Repository operators should provide additional services to facilitate sharing and usage of data e.g. pseudonymization, de-identification, anonymization, record linkage. 4) The effort and benefit of FAIR verification of data is generally unevenly distributed at the expense of the data owner. However, an incentive for them may be to receive feedback on improved or additionally calculated data and to add it to the original data source.

4. Discussion and conclusion

The FAIR data principles can well be applied to medical data repositories, yet some restrictions will apply, mostly with regard to accessibility/data privacy. The LHA project is work in progress; not all of our recommendations are solved, implemented and evaluated yet. Even through some of our findings lack a broad, independent evaluation, we expect most of them to be translatable to similar projects. A first draft of a guideline for implementing a general FAIR open data policy in health research was developed as part of the FAIR4Health project [12].

The greatest difficulties for FAIRification currently lie in the lack of availability of community-consented vocabularies and powerful tools for transforming data to common data models like HL7 FHIR. While several initiatives (e.g. FORCE11, GOFAIR, Research Data Alliance) assess the coverage of the FAIR principles of different systems and publish FAIR metrics onto how these principles should be put into practice, effort and costs are currently too high for many common data collection projects. Incentives are needed to take on the challenge of FAIRification. Furthermore, there should be more standardization on processes for data access and data extraction.

Acknowledgements: This work was supported by the European Union (H2020/824666 – FAIR4Health), the German Ministry of Education and Research (LHA: 031L0026, SMITH: 01ZZ1803A) and the German Research Foundation (NMDR: WI 1605/10-1).

References

[1] M.D. Wilkinson, M. Dumontier, I.J.J. Aalbersberg et.al., The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3 (2016), 160018.

[2] A. Dunning, M. de Smaele, and J. Böhmer, Are the FAIR Data Principles fair? *IJDC 12 (2017)*, 177–195.

[3] F.A. Meineke, M. Löbe, and S. Stäubert, Introducing Technical Aspects of Research Data Management in the Leipzig Health Atlas. *Stud Health Technol Inform 247 (2018)*, 426–430.

[4] K. Wolstencroft, S. Owen, O. Krebs et.al., SEEK: a systems biology data and model management platform. *BMC Syst Biol 9 (2015)*, 33.

[5] S.J. Steindel, OIDs: how can I express you? Let me count the ways. *J Am Med Inform Assoc 17 (2010)*, 144–147.

[6] J.A. McMurry, N. Juty, N. Blomberg et.al., Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol 15 (2017)*.

[7] M. Löbe, User Expectations of Metadata Repositories for Clinical Research. *Stud Health Technol Inform 253 (2018)*, 60–64.

[8] M.L. Braunstein, Healthcare in the Age of Interoperability: The Promise of Fast Healthcare Interoperability Resources. *IEEE Pulse 9 (2018)*, 24–27.

[9] V. Huser, C. Sastry, M. Breymaier et.al., Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). *J Biomed Inform 57 (2015)*, 88–99.

[10] G. Hripcsak, J. Duke, N. Shah et. al., Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform 216 (2015)*, 574–578.

[11] B.D. Athey, M. Braxenthaler, M. Haas et.al., tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Jt Summits Transl Sci Proc 2013 (2013)*, 6–8.

[12] T. Hernández-Pérez, E. Méndez Rodríguez: D2.3. Guidelines for implementing FAIR Open Data policy in health research. Available at: https://www.fair4health.eu/en/resources/project-deliverable