FAIR Data Management to Access Patient Data

Núria Queralt-Rosinach[1][0000−0003−0169−8159], Rajaram Kaliyaperumal[1][0000−0002−1215−167X], César Bernabé[1][0000−0003−1795−5930], Qinqin Long[1][0000−0002−6844−0067], Henk Jan van der Wijk[1][0000−0002−0317−110X], Barend Mons[1][0000−0003−3934−0072], and Marco Roos[1,2][0000−0002−8691−772X]

[1] Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands
[2] m.roos@lumc.nl

Abstract. The COVID-19 pandemic is challenging healthcare systems and research worldwide. Clinical observations in hospitalised patients are not ready to use efficiently and timely neither by humans nor by machines. The Leiden University Medical Center (LUMC) clinicians and researchers have united to adapt the Research Data Management (RDM) in the hospital in order to make patient data more Findable, Accessible, Interoperable and Reusable (FAIR) for humans and machines. In this paper, we present our FAIRification plan for data management to transform COVID-19 observational patient data collected in the hospital into FAIR data. Our work demonstrates that a FAIR RDM based on open Science, Semantic Web technologies, and FAIR Data Points (FDPs) is providing data infrastructure in the clinics for machine-actionable FAIR research objects that are ready to reuse and linkable to Linked Open Data (LOD).

Keywords: FAIR Data · Research Data Management · Open Science · Semantic Web · FAIR Data Point· COVID-19.

1 Introduction

The COVID-19 pandemic is challenging healthcare systems and research worldwide. Medical doctors and researchers need to answer questions to understand and rationalize personalized treatments and interventions, but clinical observations of hospitalised patients are not ready to use efficiently and timely neither by humans nor by machines. The Leiden University Medical Center (LUMC) is a research hospital in the Netherlands that encompasses a set of clinical and research groups with a unique expertise combination on immunology, biomedicine and data science. From the time of admission at the LUMC, COVID-19 patients are monitored, thereby collecting different types of data encompassing clinical observations, laboratory measurements, and various omics such as transcriptomics and metabolomics. However, these data are electronically captured in different formats. This poses a technical hurdle to seamlessly querying and analysing data. Better data infrastructure is needed to make patient data ready to use for answering patient management and clinical research questions across different patient-related datasets, possibly combining them with open science resources such as LOD.

The application of the FAIR principles [17] to (meta)data and services seeks to produce machine-actionable objects that maximise efficient and reproducible research. As initiator and one of the founders of the FAIR principles, the BioSemantics group, a computational knowledge discovery group in the LUMC, is working with domain experts, data management groups, and the IT department of the LUMC to transform COVID-19 observational patient data to FAIR digital objects, ready to be used by machines, and linkable to open available knowledge and other FAIR data. Our approach was to develop a RDM plan to improve FAIRness of the data. Here, we present a short paper of our first results and work in progress describing the solutions we applied to the LUMC data management. In the next sections we present methods used, then first results followed with a discussion and conclusion.

2 Methods

2.1 Developing a FAIR Research Data Management Plan

Identifying a Data Management Goal The first step was to determine the goal for the data management. The LUMC medical doctors have pressing questions at point of care such as ‘what are the clinical parameters that can predict the disease phase/course of a patient?’, ‘what are the biological pathways underlying patient symptoms?’, ‘how could they be positively or adversely affected by a particular treatment?’. To answer these complex questions, data needs to be integrated in a systems medicine approach, combined with external biomedical knowledge, and ready for computational analysis. Ideally, analysis can be easily expanded to COVID-19 data in other hospitals. These medical questions guided the development of a FAIR plan that prioritize findability and interoperability of patient data.

LUMC Data Management From admission date until discharge, patient data are collected by different departments. The types of data are demographics information, clinical information, laboratory measurements, transcriptomics (RNA-Seq) data, metabolomics data, and if the patient is transferred to Intensive Care Units (ICU) then data related to ICU outcome. The format depends on the different Electronic Data Capture (EDC) systems used. Within LUMC, clinical and preclinical information are collected in HiX [7] and Castor EDCs [1], whereas ICU data is managed by the MetaVision software [9]. These EDC systems have different data access interfaces and use different technologies. To provide a single point of data access, research data are combined in the Opal data warehousing system. Opal is the OBiBa’s (Open Source Software for Epidemiology) core database application to store data in central data repositories that integrate under a uniform interface data collected from multiple sources, and it provides tools to import, transform and describe data [11]. Data in Opal is published in the Web through the Mica software application. Mica is a software application used to create Web data portals for large-scale studies or multiple-study consortia. It provides a structured description of consortia, study catalogs and datasets, annotated and searchable data dictionaries, and data access request management. It is built upon a multitier architecture consisting of a RESTful application server for data management and administration, and clients to create and display data in the Web [10].

FAIRness Analysis of LUMC Data To improve the Findability, Accessibility, Interoperability, and Reuse of digital assets, we performed a FAIRness analysis of (meta)data, i.e. an analysis of the FAIR status of data and metadata. We analysed data and databases in order to evaluate the FAIRification effort needed [12]. We started by analysing the observational clinical measurements raw data collected from the laboratories. Then, we analysed the databases where these data are stored, which are first in Castor databases since this is the primary EDC system used in the LUMC, second in Opal data warehouse since this system is used to integrate and store data from the various data sources. We investigated the representation (structure and format) and meaning (semantics) of the data, and the tools and technologies of each database system in order to optimize the FAIRification process of data.

FAIR Research Data Management Plan We designed a *FAIR* RDM plan (see Fig. 1) in collaboration with data providers and users for provision of digital objects linkable with open external knowledge for analysis. The FAIR RDM plan's main steps are the development of a central linking data model to enable interoperability across multiple data resources, the implementation of FDPs [15] to ensure that resources can be found through the exposure of FAIR metadata, and the setup of infrastructure for accessing data via the FDPs in analysis workflows.

2.2 Improving I in FAIR with Semantic Web Technologies and a Data Linking Model

With the goals to answer research questions of medical doctors and make clinical data interoperable and linkable with LOD, we designed semantic data models to represent knowledge based on Semantic Web technologies such as the W3C Resource Description Framework (RDF) and the Web Ontology Language (OWL). Our approach was to define a conceptual model as an abstract and reusable model to capture as much of clinical data (measurements and observations), by using standard common schemas and established ontologies and vocabularies widely-used by the biomedical community such as ones in the Open Biological and Biomedical Ontology (OBO) Foundry [16]. With this approach we created a semantic model for cytokines laboratory measurements, which are metabolites used in the clinics to monitor patient immunoresponse.

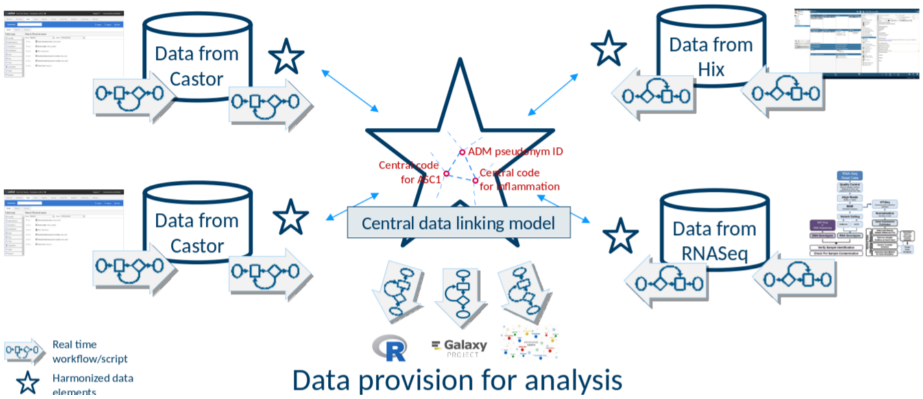

Fig. 1. FAIR Research Data Management plan. It illustrates the central concepts of our strategy: a data linking model for interoperability and FDPs for findability.

2.3 Improving F in FAIR with Semantic Web Technologies and FAIR Data Points

We implemented FDPs to make LUMC COVID-19 research objects findable for machines on the Internet. A FDP is a web application that enables data owners to expose information about their datasets using rich machine-actionable metadata. It allows creating, storing, and serving FAIR metadata about datasets and its distributions for both humans and machines. A FDP does *not* enable open access, but the metadata is expected to include information about what the resource contains and how datasets and content can be accessed under defined conditions. Opening up FAIR (meta)data by publishing them on a FDP allows algorithms to search these (meta)data, looking for patterns [6]. Mica is a tool to expose datasets from an Opal database on the Internet through Web portals that allow (meta)data descriptions. A FDP provides additional means to expose FAIR metadata, i.e. machine-actionable, via the FDP specification, a standardized metadata semantic model based on the Data Catalog Vocabulary (DCAT) [3]. FDP also exposes (meta)data via RESTful Web API that enables client applications to automate retrieval, aggregation and filtering (meta)data from distributed FDPs.

2.4 FAIR Data Analysis with Semantic Web Technologies

We used the W3C SPARQL query language to perform data analytics over the LUMC RDF patient data and across diverse external data sources in LOD. We used Blazegraph Triple Store for our use case where the data is natively stored as RDF.

Data and Code Availability The RDF semantic model and the SPARQL queries are freely available at the BioSemantics (GitHub). RDF data are accessible through the LUMC Beat-COVID FDP and queryable through the Beat-COVID Triple Store. Source code for FDP implementation is freely available at the FAIRDataPoint (GitHub).

3 Results

3.1 FAIR Status of LUMC Data Needed Improvement

We assessed that the FAIR status of LUMC data needs improvement in interoperability and findability of (meta)data on the Internet. The raw data representation from the laboratories is not uniform. The data manager then manually pre-processes and introduces all the lab information into Castor databases, which are formatted in a uniform way but without the use of common standards by default. The meaning of the data are barely described further in the Castor databases nor in the raw data files. Opal is a generic system intended for data integration, transformation and supports annotation with a vocabulary chosen by the user. The modeling it provides is for data structures and can use models that comply to for example SNOMED. It also provides a research infrastructure supporting systematic data processing and analysis. It is based on standards and it uses many open source standards for the infrastructure such as Linux, MongoDB, REST web API, or R. Opal and Mica do not directly provide semantic modeling functionality. However, Opal and Mica provide a basic annotation functionality, that is the basis of these semantic models. Opal provides annotation on the data level. Mica provides annotation on the dataset level such as how, when, where, by whom, under what conditions data has been collected. This information helps to know how data can be used in analysis and modeling, and is very valuable in making (meta)data FAIR for machines. We performed an automated FAIR analysis of a dataset described in Mica, and the results can be found here. The results show that there can be improvements made on various aspects of the metadata descriptions in Mica to make it machine-actionable and standardized metadata. Although Mica implements unique identifiers, these are not persistent and they are not explicitly defined in the metadata. This creates challenges for data accessibility and reusability. Metadata is not structured enough and grounded by ontologies, such that data is not searchable in major search engines, posing great hurdles to its findability. These challenges could be improved by implementing persistent URIs, standard metadata schemas, and describing (meta)data identifiers in the metadata. Consequently, not only would data become more FAIR but more integratable with LOD. Discussions are ongoing to manage data in a more semantic, automatic and efficient way.

3.2 FAIR Research Data Management is a Coordinated Effort

The development and execution of the RDM plan in the LUMC is a coordinated effort that, in our experience, requires at least data producers, data consumers, and data modelers who are experienced in FAIRification. This is because capturing the meaning of data requires the combined expertise of these stakeholders.

From data producers in the laboratory, data managers in the IT department, to data consumers like medical doctors, bioinformaticians, data scientists, and clinical researchers, have been involved to establish well-defined user needs, technical requirements, procedures and best practices, and to coordinate the management of the data lifecycle needs of the LUMC. To improve interoperability of data, FAIR experts have developed a semantic model for data harmonization and integration in close communication with data collectors, data managers, data analysts and medical doctors as domain experts with the real driving user needs. Whereas to improve findability of data, FAIR experts have worked with IT and database managers to develop machine-actionable metadata. Both tasks have been performed in parallel and in a synergistic way in order to support consistently the entire data management lifecycle for data analysis, and are ongoing. The whole team is maintaining one-hour bi-weekly video calls for general update and logistic discussions. For FAIRification, punctual video calls are set-up with the required set of participants and duration time depending on the topic needed to discuss for the progress. These regular and iterative meetings with all data stakeholders are necessary to enable the development of optimal semantic modeling and computational standardization since, in comparison, these latter take more time to be done and implemented. The greatest difficulties for FAIRification are 'social': the multi-disciplinar backgrounds of the different people involved. That the meetings are virtual due to the pandemic makes communication challenging. However, we store all the material presented during the meetings in addition of recording the meetings themselves to mitigate the communication gap. Due to the urgency of the COVID-19 pandemic, a single source of funding was lacking for the collaboration, which posed an extra hurdle to coordinate and keep up the FAIRification work. A big issue for data analysis is the access to real world observational patient data due to data privacy. The setup of the governance of health data for research is a bottleneck for the project.

3.3 Semantic Models for Interoperability of Clinical Measurements: Cytokines

To create a user-centered research-driven data infrastructure, we used the medical research questions as drivers for the data modeling. We first created a general concept model for the questions to extend with relevant clinical data. When we received the first actual data, cytokines clinical measurements, we created an RDF semantic model for this data (see Fig. 2). The cytokines model is based on the core module of the semantic model that was developed in the European Joint Programme Rare Diseases (EJP RD) [5] for common data elements in rare disease patient registries. This is a simple model that abstracts that every element in a patient registry is the outcome of a process, so that **process** becomes the core concept of the model [4]. We reused this model jointly with the quantitative trait semantic model [14] to capture clinical data measurements, where the process of measurement is the core concept. Reusing these existing semantic models for observational data in the LUMC supports FAIR data. Not only does it allow interoperability with patient registries and quantitative traits, but also the common biomedical ontologies used allow data integration with external knowledge such as LOD.

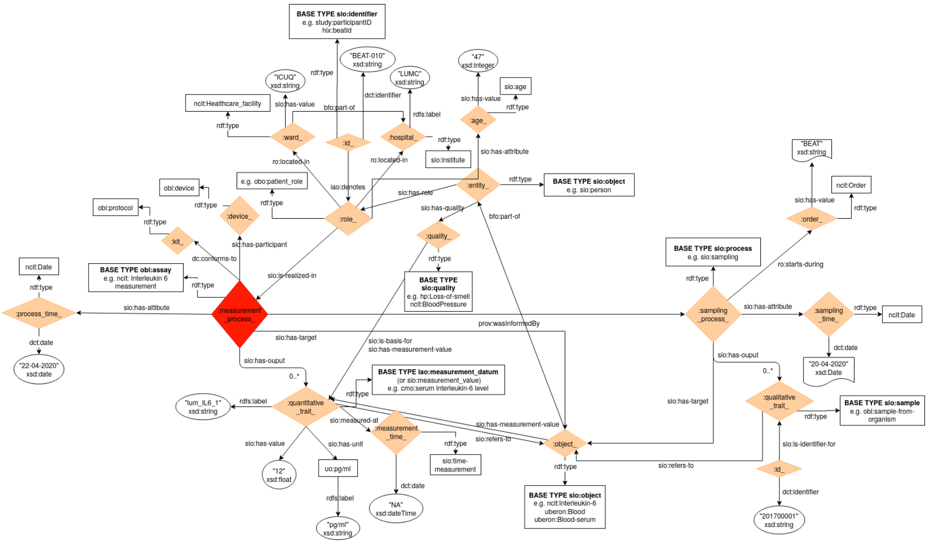

Fig. 2. Cytokines RDF semantic model. The 'process of measurement' core concept of the model is highlighted in red.

3.4 LUMC FAIR Data Point for Findability of Patient Data

We implemented a FDP to describe datasets in Opal and to publish FAIR metadata of these datasets on the Internet as complementary to the Mica system. FDP publishes structured metadata for machines to interpret how to access and use the data stored in Opal, for instance to those algorithms visiting the data with the right access. Important to the FDP approach is that the data never leave its repository thereby protecting patient data and ensuring only authorized users have access. We performed an automated FAIR analysis of one of the datasets described in the FDP. The results can be found here. In contrast to the Mica analysis results, FDP analysis results show that various aspects of the metadata description are improved when the same dataset is described in the FDP.

3.5 Querying FAIR Patient Data with LOD for Medical Questions

To showcase that the FAIR RDM and the derived data infrastructure allow querying patient data with external open science knowledge for medical questions, we developed some simple SPARQL queries performed on synthetically generated data accessible on GitHub. The queries were defined to answer the initial real-world medical doctors hypothesis related to cytokines FAIR data. For instance the user can retrieve clinical information from the LUMC such as ”count number of patients” (LUMC query), and can cross LUMC clinical data with biomedical databases such as ”retrieve measured cytokines in the LUMC with protein annotation from the UniProt knowledgebase” (federated query).

4 Discussion and conclusion

To the best of our knowledge we provide the first FAIR RDM plan for FAIRifying medical data in the hospital, that encompasses FAIR research objects ready to be used for analysis and application, and linkable to LOD. The greatest difficulties for FAIRification of the RDM in a healthcare organization were ‘social’, presumably because stakeholders are not familiar with the steps that are needed to make a resource reusable for computers outside of an organisation and one experiment. Communication is challenging and the pandemic poses an extra hurdle. For similar reasons, putting a FAIR open data policy in place for health research data conform [2] is not a usual step. Underdeveloped metadata on accesibility and data privacy hampers interoperability outside of the LUMC. Also, very important for accessibility/data privacy is that the digital objects can accommodate the criteria and protocols necessary to comply with regulatory and governance frameworks. Further, the lack of tools and vocabularies for transforming data to common data models like HL7 FHIR [8] is another big technological issue [13]. However, the use of an abstract Semantic Web data model like the EJP RD core model in combination with the implementation of FDPs may facilitate overcoming these issues. Given the urgency of the COVID-19 pandemic, we present our first results about FAIRifying data management and transforming COVID-19 observational patient data into FAIR research objects that are ready to reuse. Our work demonstrates that a FAIR RDM plan based on open Science, Semantic Web technologies, and FDPs is providing data infrastructure in the clinics for FAIR research objects linkable to LOD for analysis.

Acknowledgments

We would like to specially thank to our Beat-COVID colleagues in the LUMC: Sesmu Arbous, Jacqueline Janse, Erik Flikkenschild, Simone Joosten, Karin van der Pal, Bram Wesselo, Hailiang Mei, Annika Jacobsen, Eleni Mina, Tooba Abassi-Daloii, Daniël Wijnbergen, Winette Koning, Kees Burger, Luiz Olavo Bonino da Silva and Katy Wolstencroft. Also to our colleague Peter-Bram’t Hoen and our EJP RD colleague Mark Wilkinson for all the discussions. N. Queralt-Rosinach, R. Kaliyaperumal, C. Bernabé, Q. Long and M. Roos are supported by funding from the European Union’s Horizon 2020 research and innovation program under the EJP RD COFUND-EJP N° 825575. We would also like to thank to the EJPRD, the GO FAIR VODAN, and the ZonMW Health Holland under the Trusted World of Corona, for supporting the research on FAIR data stewardship that was reused here. Finally, we would like to thank to Professor Barend Mons for inspiring us to make a real difference in data sharing and knowledge representation.

References

1. Castor Homepage. https://www.castoredc.com/, last accessed 2020/08/20
2. D2.3. Guidelines for implementing FAIR Open Data policy in health research. https://www.fair4health.eu/en/resources/project-deliverable, last accessed 2020/08/24
3. DCAT2 W3C Homepage. https://www.w3.org/TR/vocab-dcat-2/, last accessed 2020/08/24
4. EJP RD core semantic model Homepage. https://github.com/ejp-rd-vp/CDE-semantic-model/wiki/Core-model-SIO, last accessed 2020/10/18
5. EJP RD Homepage. https://www.ejprarediseases.org/, last accessed 2020/08/24
6. FDP specification Homepage. https://github.com/FAIRDataTeam/FAIRDataPoint-Spec, last accessed 2020/08/24
7. HiX Homepage. https://www.chipsoft.com/solutions/550/Solutions, last accessed 2020/10/18
8. HL7 FHIR Homepage. https://www.hl7.org/fhir/, last accessed 2020/10/18
9. MetaVision iMDsoft Homepage. https://www.imd-soft.com/products/intensive-care, last accessed 2020/08/20
10. Mica OBiBa's software Homepage. https://www.obiba.org/pages/products/mica/, last accessed 2020/08/20
11. Opal OBiBa's software Homepage. https://www.obiba.org/pages/products/opal/, last accessed 2020/08/20
12. Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L.O.B., Mons, B., Schultes, E., Roos, M., Thompson, M.: A generic workflow for the data fairification process. Data Intelligence **2**(1-2), 56–65 (2020). https://doi.org/10.1162/dint_a_00028, https://doi.org/10.1162/dint_a_00028
13. Löbe, M., Matthies, F., Stäubert, S., Meineke, F.A., Winter, A.: Problems in fairifying medical datasets. In: Pape-Haugaard, L.B., Lovis, C., Madsen, I.C., Weber, P., Nielsen, P.H., Scott, P. (eds.) MIE. Studies in Health Technology and Informatics, vol. 270, pp. 392–396. IOS Press (2020), http://dblp.uni-trier.de/db/conf/mie/mie2020.html#LobeMSMW20, conference cancelled because of Covid-19.
14. Queralt-Rosinach, N., Bello, S., Hoehndorf, R., Weiland, C., Rocca-Serra, P., Schofield, P.N.: Modeling quantitative traits for covid-19 case reports. medRxiv (2020). https://doi.org/10.1101/2020.06.18.20135103, https://www.medrxiv.org/content/early/2020/06/20/2020.06.18.20135103
15. da Silva Santos, L.O.B., Wilkinson, M.D., Kuzniar, A., Kaliyaperumal, R., Thompson, M., Dumontier, M., Burger, K.: Fair data points supporting big data interoperability. London: ISTE Press pp. 270–279
16. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. Nature Biotechnology **25**(11), 1251–1255 (November 2007). https://doi.org/doi:10.1038/nbt1346

17. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific data **3** (2016)