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

doi:10.3233/SHTI210281

FAIRifying Clinical Studies Metadata: A Registry for the Biomedical Research

Vittorio MELONI[a,1], Alessandro SULIS[a], Cecilia MASCIA[a] and Francesca FREXIA[a]
[a] *CRS4: Center for Advanced Studies, Research and Development in Sardinia, Italy*

Abstract. The data produced during a research project are too often collected for the sole purpose of the study, therefore hindering profitable reuse in similar contexts. The growing need to counteract this trend has recently led to the formalization of the FAIR principles that aim to make (meta)data Findable, Accessible, Interoperable and Reusable, for humans and machines. Since their introduction, efforts are ongoing to encourage FAIR principles adoption and to implement solutions based on them. This paper reports on the FAIR-compliant registry we developed to collect and serve metadata describing clinical trials. The design of the registry is based on the FAIR Data Point (FDP) specifications, the state-of-the-art reference for FAIRified metadata sharing. To map the metadata relevant to our use case, we have extended the DCAT-based semantic model of the FDP adopting well-established ontologies in the biomedical and clinical domain, like the Semanticscience Integrated Ontology (SIO). Current implementation is based on the Molgenis software and provides both a user interface and a REST API for metadata discovering. At present the registry is being loaded with the metadata of the 18 clinical studies included in the `I FAIR Program', a project finalised to the dissemination of FAIR best practices among the clinical researchers in Sardinia (Italy). After a testing phase, the registry will be publicly available, while the new model and the source code will be released open source.

Keywords. FAIR registry, Digital Health, Fair Data Point, Clinical Trials, Molgenis, DCAT, SIO

1. Introduction

The difficulties of using in other contexts the datasets created during a clinical trial [1], as well as the low reproducibility of the results obtained [2], have been widely analysed in order to overcome the barriers preventing an efficient reuse of quality data [3, 4]. The cataloging of clinical trials in public registries is recognised as an effective way to improve transparency and data quality [5]. Several registries have been created and are offering a very important source of information[2], but they still present limitations in the effective representation of the collected data according to standard terminologies and ontologies [6], essential factor to support data discoverability and repurposing. A promising approach to face these issues is based on the application of the FAIR

[1] Corresponding Author, Vittorio Meloni, CRS4: Center for Advanced Studies, Research and Development in Sardinia, Loc. Piscina Manna - Edificio 1, 09050 Pula (CA), Italy; Email: vittorio.meloni@crs4.it.

[2] ClinicalTrials.gov (clinicaltrials.gov), ISRCT (isrctn.com), EU Clinical Trials Register (clinicaltrialsregister.eu), NIHR Be Part Of Research (bepartofresearch.nihr.ac.uk), International Clinical Trials Registry Plat- form (www.who.int/clinical-trials-registry-platform).

Principles [7], a set of guidelines focused on the goal of making Findable, Accessible, Interoperable and Reusable all the 'research objects' (like data, algorithms and workflows) generated during a research project. The FAIR Principles aim to support information discovery, both by humans and machines, through good data generation and management. Their application is strongly encouraged in all research fields [8] and is experiencing the most significant progress in biological and natural sciences [9], but at present there are not FAIR implementations of clinical trial metadata resources.

In this work, we present the registry for biomedical research that we are developing to offer a FAIR access to significant metadata describing clinical studies. The paper is focused on the design of the semantic metadata model, the system implementation, its present status, with the initial population with the metadata of 18 clinical studies, also providing an overview about the next development steps and the future use.

2. Methods

The objective of the registry is twofold: the collection of significant information about biomedical research trials and the development of interfaces enabling humans and software agents to search and access this information, according to FAIR principles.

The design has been based on the specification of the Fair Data Point (FDP) [10], the current state-of-the-art for exposing metadata of datasets in a FAIR way, which provides a metadata model and a REST API to access the metadata. The metadata model, mainly based on Data Catalog Vocabulary (DCAT) [11], is a hierarchical structure composed of complementary layers that we mapped to the hierarchy of the registry as follows: (i) *FDP*, providing information about the FDP itself, enabling its inclusion in a federated network of FDPs; (ii) *Catalog*, containing information about a collection of Datasets, i.e. the group of clinical studies; (iii) *Dataset*, specifying information about the datasets, which in our case are the trials; (iv) *Distribution*, giving information about the specific format of the datasets. For every layer we paired our metadata to the FDP ones. In particular, for the Dataset layer, in some cases we found a natural matching (e.g., *name* - *dcat:title*; *targeted disease* - *dcat:theme*; *license of usage* - *dcat:license*). Instead, for several domain-specific concepts, the FDP had no correspondence that could fit, and it was therefore necessary to extend the model to include essential attributes, such as: the therapeutic area, the type of collected data (e.g., lab exams), the experimentation center, the ethical committee. For this scope we adopted the Semanticscience Integrated Ontology (SIO) [12].

The registry uses terms from clinical and biomedical ontologies for the values of the metadata, to enable semantic interoperability and machine-actionability, as recommended by FAIR guidelines [15]. For example, we use ICD [16] for diseases and MeSH [17] for therapeutic areas.

The implementation of the registry is based on Molgenis [18], an open-source application for modelling, collecting, managing and accessing complex datasets. This choice is motivated by some of the Molgenis main features, since it has a native implementation of the FDP, a REST API to query the data, authentication and authorization functionalities and it gives the possibility of extending the model according to our needs and implementing a custom web user interface.

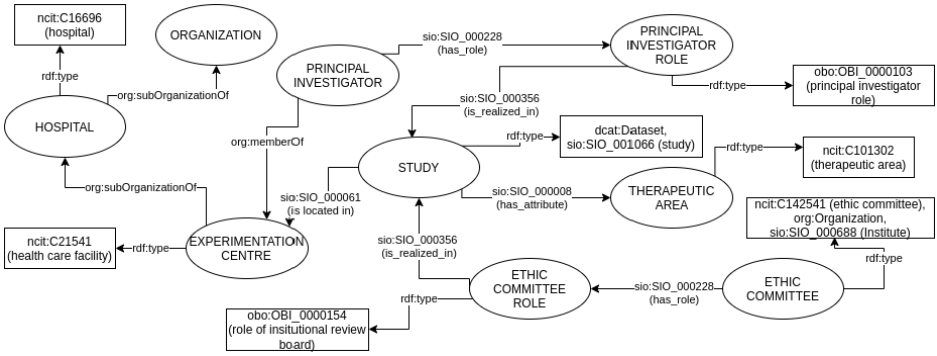

Figure 1. Extension of the FDP model. New concepts (ellipses) are connected to the Dataset/Study with SIO properties (arrows) and denoted with OBI and NCIT terms (rectangles)

3. Results

We have implemented a registry to collect and expose a set of general metadata describing clinical trials, FAIRly usable by humans and machines, starting from the FDP specifications.

To include all the relevant metadata for our use case, we extended the DCAT-based FDP semantic model adding new concepts connected to the Study by SIO ontology properties and denoted in a meaningful way using terms from specific ontologies, such as Ontology for Biomedical Investigations (OBI) [13] and National Cancer Institute Thesaurus (NCIT) [14]. *Principal Investigator* and *Therapeutic Area* are two examples in this sense and are shown in Figure 1, which illustrates the extended semantic model. The *Principal Investigator* is assigned a *Role* with the *sio:has_role* property and then the *Role* is connected to the *Study* with *sio:is_realized_in* property. The *Role* is denoted by *obo:OBI_0000103* (Principal Investigator Role in OBI). For *Therapeutic Area*, we linked the *Study* with an object of type *ncit:C101302* (Therapeutic Area in NCIT) by means of property *sio:has_attribute*. The semantic modelling activity is still in progress: all the generic attributes have already been included, while the mapping of the study-specific metadata has been partially completed and the missing ones (e.g., collected specimen type and collected data type) will be mapped in future releases. Some of the values for the metadata of the current model have been already stored into a dedicated Molgenis instance, while others require the collaboration of data stewards of the studies to be loaded (for example, the licence for the dataset, the access rights or the collected specimen type).

The metadata are available for machines through the Molgenis REST API which exposes their Turtle [19] serialization by means of the FDP module. For humans, we designed and implemented a web interface leveraging Molgenis web app, to enable searching using some predefined parameters (like disease, therapeutic area or status of the trial) and the direct access to the details of trials, including the study protocol and the ethical documentation. At present, the registry includes metadata of the 18 studies of the 'I FAIR Program'[3], a regional project to promote the FAIR culture among clinical researchers in Sardinia (Italy) by combining educational activities and financial

[3] https://www.sardegnaricerche.it/index.php?xsl=370&s=410331&v=2&c=6072&nc=1&sc=

support for data stewardship. After a testing phase, the registry will be accessible to the public and the model together with the source code will be made available open-source in order to make them reusable.

4. Discussion

This work presents the implementation of a FAIRified registry enabling the sharing of metadata about clinical studies, in order to enhance the potential reuse of the data acquired during the trials' execution. Despite the increasing popularity of the FAIR Principles, our registry is, to our knowledge, among the first examples in this context.

In our work, we refer to the FDP specifications for both semantic metadata model and implementation, a choice consistent with the recommendation of using existing models for FAIRification [15]. The FDP model offers a robust set of generic metadata to describe datasets and a hierarchical structure that allows inclusion of the registry into federated networks. On the other hand, when starting from the general-purpose design of the FDP, it's necessary to define a custom strategy to support use case specific needs, like the exposition of domain-related metadata about the clinical studies. Our approach to this task is based on the adoption of a unique high-level ontology (in our case SIO) with generic properties that can be exploited to connect the concepts. Every concept is given unambiguous meaning, associating it to a term taken from a specific ontology (as NCIT and OBI for the biomedical domain). The number of these ontologies should be limited as much as possible. A promising alternative for the high-level ontology is DCAT 2, recently published but still not adopted by the FDP specification. This new version provides a larger set of properties and it would be interesting to evaluate its adoption in place of SIO, as it would streamline the modelling by avoiding one dependency.

On the implementation side, our work is based on Molgenis as it is open-source, which allows customizations, and it provides native FDP support, an easy way to extend and populate the semantic model, a good ontology support and a consolidated user base. Other solutions have emerged since the start of our work and can be evaluated: one of the most promising is the FDP reference implementation, because, being based on semantic web technologies, it could facilitate semantic interoperability and *ontologized* querying.

5. Conclusions and Future Developments

In this work we presented the implementation of a registry enabling to expose and share a set of relevant attributes about biomedical studies. The present implementation leverages Molgenis functionalities in terms of FAIR support, authentication and data access, also extending its FAIR Data Point model to map the studies' metadata, which are made available via interfaces both for humans and machines.

Future developments include: (i) the finalisation of the metadata model with all the study-specific metadata identified in collaboration with the researchers for each of the study in the 'I FAIR Program'; (ii) the evaluation of the complete version of the registry with the metrics [20] developed by the FAIR community; (iii) the implementation of semantic web queries support, to further improve data discoverability; (iv) the extension of the user interface with new filters.

The registry and the extended FDP data model will be released in open-source and will be available for reuse in other contexts.

Acknowledgements

This work has been partially supported by the FAIR_DATA Project (funded by Sardegna Ricerche within the 'I FAIR Program', FESR 2014/2020), the DIFRA Project (funded by the Sardinian Regional Authority) and by the European Joint Programme on Rare Disease (grant agreement N. 825575).

References

[1] Vines TH, Albert AY, et al. The availability of research data declines rapidly with article age. Current biology. 2014;24(1):94–7.
[2] Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol. 2015 Jun 9;13(6):e1002165.
[3] Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open. 2017;7:e018647.
[4] Wilkinson T, Sinha S, Peek N, Geifman N. Clinical trial data reuse - overcoming complexities in trial design and data sharing. Trials. 2019;20(513).
[5] Al-Shahi Salman R, Beller E, et al. Increasing value and reducing waste in biomedical research regulation and management. Lancet. 2014 Jan 11;383(9912):176-85.
[6] Miron L, et al. Obstacles to the reuse of study metadata in ClinicalTrials.gov. Sci Data. 2020 Dec 18;7(1):443.
[7] Wilkinson MD, Dumontier M, Aalbersberg JI, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018.
[8] Directorate-General for Research and Innovation (European Commission). Turning FAIR into reality. Final Report and Action Plan From The European Commission Expert Group on FAIR Data. Research and Innovation. 2018.
[9] van Reisen M, Stokmans M, Basajja M, Ong'ayo AO, Kirkpatrick C, Mons B. Towards the Tipping Point for FAIR Implementation. Data Intelligence. 2020;2:1-2:264-275.
[10] da Silva Santos LOB, Wilkinson MD, et al. FAIR Data Points Supporting Big Data Interoperability [Internet]. In: Zelm M, Doumeingts G and Mendonça JP. Enterprise Interoperability in the Digitized and Networked Factory of the Future [Internet]. iSTE Press. 2016:270–279.
[11] Maali F, Erickson J, Archer P. Data Catalog Vocabulary (DCAT). The World Wide Web Consortium. 2014.
[12] Dumontier M, Baker CJO, Baran J, et al. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. J Biomed Semant. 2014 Mar 06;5(1):14.
[13] Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The Ontology for Biomedical Investigations. PLoS ONE. 2016 Apr;11(4):e0154556.
[14] Sioutos N, de Coronado S, et al. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007 Feb;40(1):30-43.
[15] Jacobsen A, Kaliyaperumal R, da Silva Santos LOB, Mons B, Schultes E, Roos M, Thompson M. A generic workflow for the data FAIRification process. Data Intelligence. 2020;2(1-2):56-65.
[16] World Health Organization (WHO). The ICD-10 classification of mental and behavioural disorders. World Health Organization. 1993.
[17] Rogers FB. Medical subject headings. Bull Med Libr Assoc. 1963 Jan;51(1):114-6.
[18] Van der Velde KJ, Imhann F, Charbon B, Pang C, et al. MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians. Bioinformatics. 2019 Mar 15;35(6):1076-8.
[19] Prud'hommeaux E, Carothers G. RDF 1.1 Turtle. W3C Recommendation. 2014 Feb. Available at: https://www.w3.org/TR/turtle/.
[20] Wilkinson MD, Dumontier M, Sansone SA, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. Sci Data. 2019 Sep 20; 6(1):174.