Enabling FAIR Clinical Data Standards with Linked Data

Javier D. Fernández, Nelia Lasierra, Didier Clement, Huw Mason, and Ivan Robinson

Medical Data and Information Solutions, F. Hoffmann-La Roche, Basel, Switzerland
name.surname@roche.com

Abstract. This article reports on our efforts to support FAIR Clinical Data Standards with Semantic Web technologies, including the challenge of bridging the gap for non-technical users.

1 Introduction

The biopharmaceutical industry is traditionally led by strict (clinical) standards to regulate how clinical trial data are collected, tabulated, analyzed, and finally submitted to regulatory authorities. However, with the advent of the data deluge era, adherence to data standards not only ensures data will meet regulatory expectations, but it can also spark and fuel scientific insights when mastering well-curated, integrated, and complex data.

In this context, the novel concepts of FAIR data [3], and the corresponding FAIRification processes play a crucial role. FAIR provides guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. FAIR emphasises machine-actionability and data-driven processes to deal with the current increase in volume, complexity, and creation speed of data. The biopharmaceutical industry and academia have rapidly embraced these principles to improve its efficiency [4]. Proof of that is the collaboration emerging within the non-profit Pistoia Alliance[1], that pursuits these efforts.

The Roche Global Data Standards Repository (GDSR) is a system that stores and retrieves selected data for different information domains in the Roche product development system landscape. Although GDSR does not restrict the domain of the data, the main source of information has traditionally been Clinical Global Data Standards (GDS), aligned with CDISC[2] standards.

In the following, we show how GDSR adheres to FAIR principles thanks to the underlying Linked Data technology, and how these concepts are brought to non-technical users.

[1] http://www.pistoiaalliance.org
[2] https://www.cdisc.org/

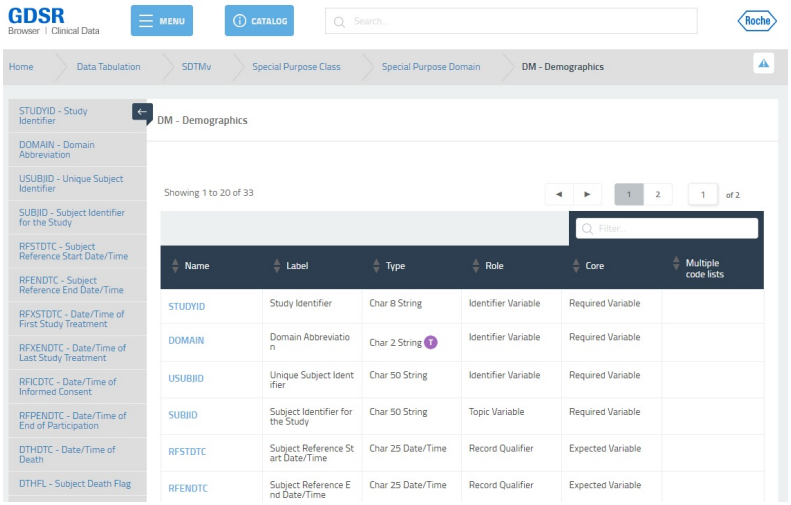


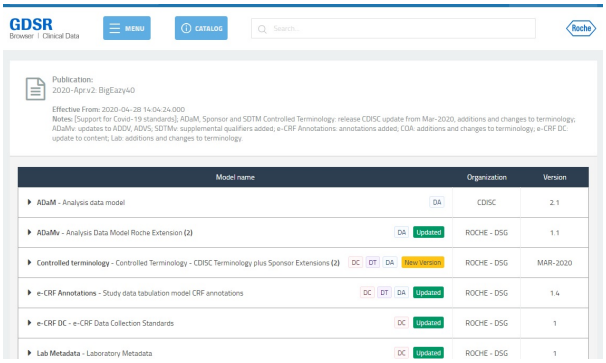
Fig. 1: The GDSR UI to browse Linked Data by non-technical users.

2 Enabling FAIR Clinical Data Standards

GDSR includes all the required standards (and extensions) to ensure consistency across our clinical trials. These standards are kept in GDSR in semantic graphs: clinical data standards are modeled as OWL/RDF ontologies and vocabularies, conforming a knowledge graph for (meta) data assets, which ensure common understanding and facilitates integration and sharing. In the following, we describe how GDSR supports FAIR principles.

2.1 Findable

Following Linked Data, each concept in GDSR is identified and can be referenced using URIs, e.g. http://gdsr.roche.com/instrument-lab#Analyte.ADIPOCYTE. The semantic data are indexed in a triplestore to support searching capabilities via a human-friendly GDSR browser for non-technical users (see Figure 1), or through pre-defined SPARQL queries. In turn, metadata about the standards is provided by a summary catalog (see Figure 2a), together with diverse predefined human-readable reports. We also make intensive use of RDF versioning, so URIs can be found across any of the GDSR archives where it is present. To do so, we implement a independent copy approach [1], maintaining the full copy of the data in different time snapshots, called publications. In each publication cycle, the new, updated, and retired data standards are made available to users and the previous publication is moved to an accessible archive (see Figure 2b).



(a) GDSR Catalog

(b) Accessing versions

Fig. 2: GDSR catalog and access to different (graph) versions in the GDSR UI.

2.2 Accessible

Besides the aforementioned GDSR browser, GDSR leverages Linked Data to make content available through a public REST API, which is also bundled in an R package, called rGDSR, to access the information programmatically. This allows non-semantic web experts to still access the full data standard catalog programmatically and to create automated processes. In particular, rGDSR is used to maximize the use of data standards and terminologies in clinical studies during study design, as well as to detect potential deviation to standards in the study. It is worth noting that, due to the aforementioned gap, only technical users have direct access to the internal SPARQL endpoint while results of pre-defined SPARQL queries are served to all users.

2.3 Interoperable

Using Linked Data, GDSR supports the sharing and use of metadata with other semantic technologies and systems across Roche, such as the Roche Terminology System (RTS) [2]. In addition, GDSR links to external datasets such as CDISC ontologies, enabling the exploration of connected vocabularies. In fact, our ongoing work focuses on leveraging federated SPARQL queries to access and combine data available in multiple and disparate semantic systems at Roche, enabling seamless access across that ecosystem.

2.4 Reusable

Clinical data standards stored in the GDSR as Linked Data are accessible and valid globally and across molecules, study phases and therapeutic areas. We make use of the PROV-O ontology for managing changes in the Knowledge Graph from one publication to another, including the change description, owner and responsible entities. This information is also presented in a human-friendly version and via the aforementioned REST API to bridge the gap for non-technical users.

3 Conclusions

The biopharmaceutical industry, and Roche in particular, is embracing the FAIR principles to improve data-driven efficiency, leading to novel findings and faster filings, which means faster access to novel medicines for patients. This work shows how FAIR principles are translated to clinical data standard management in practice, thanks to Linked Data technologies.

Our ongoing work regards the challenge of authoring and managing existing clinical data standards, in the form of Linked Data, by non-semantic web experts. Thus, we are working on a visual editor to facilitate this task while still assuring the expected quality and conformance of the standards. To do so, we plan to leverage the SHACL W3C standard to represent and enforce key quality aspects of the data.

Acknowledgments

This work was been possible thanks to the support of the DAAV Information Architects, the GDSR dev team and the RGITSC Custom Apps team.

References

1. Fernández, J.D., Umbrich, J., Polleres, A., Knuth, M.: Evaluating Query and Storage Strategies for RDF Archives. Semantic Web **10**(2), 247–291 (2019)
2. Thalhammer, A., Romacker, M., Rupp, J.: Semantic terminology management for applications: Contextualized skos-xl. In: Proc. of the International Semantic Web Conference (ISWC) Posters, Demos & Industry Tracks (2017), http://ceur-ws.org/Vol-1963/paper477.pdf
3. 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)
4. Wise, J., de Barron, A.G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E., Mellino, G., Harrow, I., Smith, I., Taubert, J., et al.: Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug discovery today **24**(4), 933–938 (2019)