Paper PP06

GDSR FAIR – FAIR Data Standards in the GDSR (Global Data Standards Repository)

Huw Mason, F Hoffmann - La Roche, Basel, Switzerland

ABSTRACT

FAIR data principles support the curation and integration of clinical trial data allowing for the reuse of data and the creation of pooled data marts where enhanced scientific insights can be extracted. In this poster the application of FAIR data principles: 1) Findable 2) Accessible 3) Interoperable 4) Reusable, in a data standards repository are presented. The poster will focus on the storage, technology and FAIR aspects of the Global Data Standards Repository (GDSR) at Roche, and the importance of data standards in the clinical trial lifecycle.

INTRODUCTION

The use of data standards in the clinical trial lifecycle contribute to the I of FAIR enabling interoperability and allowing for a consistent approach to the combining of data from across sites as well as the collection, tabulation, and analysis of data. In turn the data standards support the creation of standard processes, documents, programs, macros and systems to streamline workflows with the aim of getting medicines to patients faster.

The GDSR is Roche and Genentech's single source of truth for clinical data standards, storing and making them available to users globally in the most business useful format. The data standards are stored in semantic graphs and the technology behind the GDSR is based on RDF, ontologies, and linked data principles aligned with Semantic Web (WC3) standards. Having common vocabularies for (meta) data assets ensures common understanding and facilitates integration and sharing for machines.

From the Roche FAIR Data playbook:
*"FAIR data are best stored in a system that provides unique data asset IDs, tracking and management of current and historical versions, and facilities for structured annotation with metadata. Suitable systems must support an organized set of annotation tags and metadata that describe and characterize data and permit easy inspection and querying of the metadata by individuals and by other computing systems." *1*

FAIR PRINCIPLES *2

In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in Scientific Data. The authors intended to provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

FINDABLE

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

ACCESSIBLE

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

INTEROPERABLE

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

REUSABLE

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

FINDABLE - GDSR

How does the GDSR meet the Findable principle?

UNIQUE RESOURCE IDENTIFIER (URI)

The GDSR uses RDF and ontologies to describe the metadata and standard models, hence each concept registered in the GDSR is identified using a URI.
<http://gdsr.roche.com/cdisc/adamig#Table.ADSL>
An instance of metadata can be referenced or found using its URI across any of the GDSR archives where it is present.

CATALOG

The GDSR has a catalog providing metadata about the data standard models (eCRF, Controlled Terminology, SDTM, ADaM, Non-CRF, etc.). Each publication of clinical data standards contains a unique catalog providing users with a summary of the models available in that specific publication, their descriptions, version and revision numbers.

REPORTS

The GDSR generates predefined user reports, with metadata from facets, that give detail to the components of a model.

SEARCHING

The GDSR supports indexed searching; either through the GDSR browser or using Roche's Search Across Systems application. Users can search for metadata using any of the properties of a given instance.

ACCESSIBLE - GDSR

How does the GDSR meet the Accessible principle?

REST API

The GDSR makes content available through a public REST API. Users can programmatically access the data standards, catalog, publication information, and the archives, making automation of previously manual tasks possible.

BROWSER

The GDSR makes content available through a browser allowing users to explore the standards, catalogs, and archives through a user friendly interface.

PUBLICATION VERSIONING

The GDSR uses publication versioning which allows users to access the archives of previous data standards versions. New, updated, and retired data standards are made available to users in a publication cycle and the previous publication is moved to an accessible archive.

INTEROPERABLE - GDSR

How does the GDSR meet the Interoperable principle?

ONTOLOGIES

The GDSR uses ontologies and controlled vocabularies to define the standards metadata. Metadata is created and described using RDF/OWL and internally developed ontologies meaning it can be integrated with other semantic technologies and systems.

SCHEMA

All metadata and schema in the GDSR are described in ontologies, supporting the sharing and use of metadata with other semantic technologies and systems.

LINKED DATA

Data standards described in the GDSR link to external datasets such as CDISC metadata and ontologies, enabling the exploration of connected vocabularies.

REUSABLE - GDSR

How does the GDSR meet the Reusable principle?

VAILD GLOBALLY

Clinical data standards stored in the GDSR are accessible and valid globally and across molecules, study phases and therapeutic areas. The catalog contains authorship of data standards models.

PROVENANCE

The Provenance ontology PROV-O is used to capture information about changes to the models between versions. Descriptions of changes to models, model owners, and entities responsible for changes are captured and detailed in the catalog.

COMMUNITY TRENDS

GDSR technology follows community trends, it is aligned with Semantic Web (W3) standards and vocabularies, as well as Pharma best practices (CDISC). The GDSR is flexible and stays aligned with evolving trends and regulatory requirements.

CONCLUSION

In order to generate the enhanced scientific insights that help guide the decisions that lead to the development of the medicines patients need, the vast amounts of diverse data collected across the business need to be FAIR. The FAIRification of data is one step in reducing the obstacles that obstruct the streamlining of processes and data transformation. Having FAIR systems and technologies, such as the GDSR, not only helps the business to meet regulatory requirements and take another step in the direction of reducing those obstacles, but also to adopt FAIR practices into our day to day work and encourage good data citizenship.

REFERENCES

*1 Roche FAIR DATA Playbook
*2 https://www.go-fair.org/fair-principles/.

ACKNOWLEDGMENTS

Nelia Lasierra, Didier Clement, Ivan Robinson

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Huw Mason
F Hoffmann – La Roche Ltd
124 Grenzacherstrasse
Basel 4058
Switzerland
Work Phone: +41 (0)61 68 22497
Email: huw.mason@roche.com
Web: www.linkedin.com/in/huw-mason