Enrichment of rare disease patient registries by FAIRification

Annika Jacobsen[1], Mark Thompson[1], Erik A Schultes[2], Ronald Cornet[3], Mark D Wilkinson[4], Rachel Thompson[5], Marina Mordenti[6], Luca Sangiorgi[6], Claudio Carta[7], Karlijn Groenen[8], Martijn G Kersloot[3,9], Leo S Kool[8], David van Enckevort[10], Marco Roos[1]

[1]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands, [2]GO FAIR International Support and Coordination Office, Leiden, The Netherlands, [3]Academic Medical Center, dept of Medical Informatics, Amsterdam, The Netherlands, [4]Universidad Politécnica de Madrid, Spain, [5]MRC Centre for Neuromuscular Diseases, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, The United Kingdom, [6]Medical Genetic Department & CLIBI Lab, Rizzoli Orthopaedic Institute, Bologna, Italy, [7]National Centre for Rare Diseases, Istituto Superiore di Sanità, Rome, Italy, [8]Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Nijmegen, The Netherlands, [9]Castor EDC, Amsterdam, The Netherlands, [10]University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

We acknowledge the generous support from RD stakeholders, RD-Connect, ELIXIR, ELIXIR-EXCELERATE, BBMRI-ERIC, ODEX4All and FAIR-dICT.

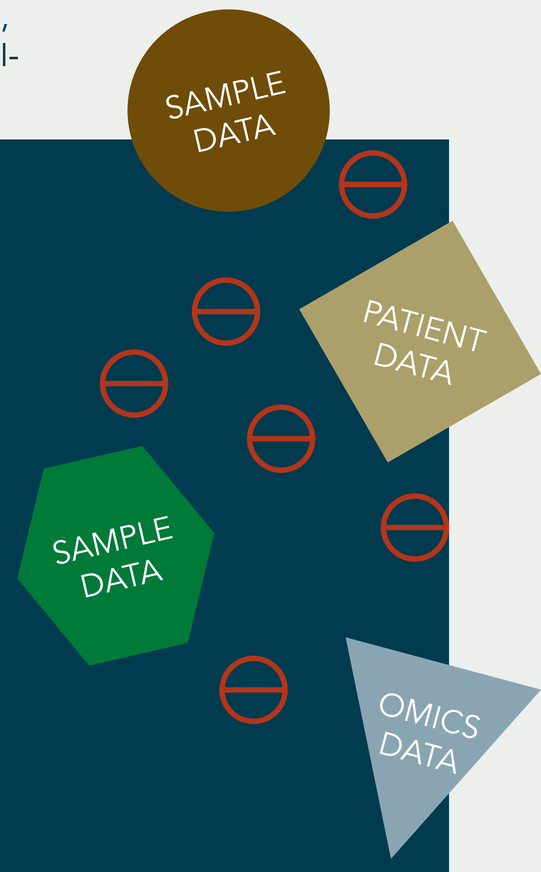
FAIRification of rare disease registries

Rare disease (RD) registries contain valuable information to advance our knowledge on diagnosis, disease progression, and treatment of rare diseases. However, these data are often sensitive, sparse, highly distributed and heterogeneous.

Making these data FAIR (Findable, Acceissible, Interoperable, and Reusable) for humans and machines will enable them to be optimally used for research.

FAIRifying a rare disease registry, means: that the registry data and metadata are made machine-readable (*interoperable*), that the metadata clearly describes how the the data can be *accessed* and *reused*, and that the metadata is *findable* by machines. This requires complete understanding of the registry data including: i) the data items/elements, ii) how it was created, and iii) the accessibility restrictions.

We are currently FAIRifying a selected set of rare diseases such as: Vascular Anomalies (VA), Duchenne Muscular Dystrophy (DMD), Rett syndrome, and Osteogenesis Imperfecta (OI). At the same time we are also defining and developing the procedures, tools, expertise and ICT infrastructure that is required for FAIRification of rare diseases.



Machine-readable data and metadata

- The degree of interoperability is greatly impacted by the choice of ontologies by which data are described.
- The Human Phenotype Ontology and the Orphanet Rare Disease Ontology are already IRDiRC recognized resources
- Data and metadata concepts and relations should be described by globally unique and persistent identifiers.
- The original data (which could be tabular) is converted to an ontology-grounded machine-readable format called RDF.

FAIR Data Point

Rare disease metadata are:

- made *findable* for machines through a FAIR data point (FDP
- describe under which conditions the data can be accessed
- contain sufficient description for data reusability.

A FDP needs to be indexed in a search engine in order to be *finable*.

Interoperable data

In order to make data interoperable i.e., machine-readable, we first define a driving user question, e.g.:
Is there a correlation between the Orpha code and age?

Figure below: Part of semantic data model describing the minimal data elements for rare disease registries.

Metadata with clearly described accessibility and reusability

Rare disease registries often contain no or little metadata Therefore metadata often needs to be created from scratch.This requires requires complete understanding of the registry data, e.g.*: How was the data created? What are the accesibility restrictions?* This requires involvment from e.g., the data owner to describe how the data was created (*reusability*), a in-house legal expert to define the *accessibility*, and a FAIR expert to make the metadata interoperable, i.e., machine readable. The creation of machinereadable metadata can be done using the DTL Metadata Editor. We use the Data Catalouge (DCAT) vocabulary to describe metadata, and are currently defining a community standard for rare disease registries that defines a way for registries to share and reuse metadata models and address crucial concerns about data access permission and data security. Example of metadata: Title, identifier, language, keywords, themes, creator, location, protocol, format, version, authors, etc.

The question is analysed and the specific data elements and items required to answer the question(s) are identified. Unique persistent identifiers are then assigned to the data. At this stage it's important to think about what the relations are between the concepts. These will be used to create (or chose) a semantic data model that will be used to 'guide' the creation of machine-readable data. Here, we use the DTL FAIRifier.