EJP RD

European Joint Programme on Rare Diseases

H2020-SC1-2018-Single-Stage-RTD
SC1-BHC-04-2018
Rare Disease European Joint Programme Cofund

Grant agreement number 825575

Del 11.1

First Ontological model of resources metadata

Organisation name of lead beneficiary for this deliverable:

Partner 76 – ELIXIR/EMBL-EBI

Contributors: AMC, BBMRI-ERIC, GUF, INSERM-Orphanet, LUMC, UMCG, UPM

Due date of deliverable: month 12

Dissemination level: Public

Table of contents

List of figures

List of tables

1. Introduction

This deliverable is the result of the ongoing work initiated in task 11.1 of the WP11, i.e. the definition of a metadata and ontological model for cataloging resources for Rare Diseases (RD) research. This work aims to improve interoperability among resources and provide a unified way to describe a range of includable resources in the centralized suite of catalogs built in Task 11.2. The deliverable describes the first release of the EJP RD Virtual Platform (VP) ontological model and metadata model. The ontology is designed to represent the core domain concepts used to describe data elements in catalogs of registries and biobanks for rare disease, and is designed to include also patient registries, biobanks and other resources in the future. The ontology provides standard vocabulary terms that can be used by individual resources to describe their metadata elements, or to map to their existing elements to a shared model. These semantic annotations will be used by the EJP RD Virtual Platform to harmonise metadata from the various resources (catalogs, registries and biobanks) and provide unified semantics for accessing and processing data.

2. Approach

We initiated the development of the model by collecting metadata elements from existing catalogs of patient registries and biobanks. The selected catalogs for this task were the Orphanet catalog of registries and biobanks[1], RD-Connect Biobank and registry finder[2], the RD-Connect Sample catalogue[3] and the ERDRI directory of registries[4]. By collecting the metadata elements used at the catalog level, it was possible to align the common elements to define an initial model.

Our approach to developing the model is based on the ISO/IEC 11179[5] (formally known as the ISO/IEC 11179 Metadata Registry (MDR) standard), an international standard for representing metadata for an organization in a metadata registry. This provides a framework for describing the metadata model in terms of concepts, relationships between concepts and attributes/characteristics that capture specific data values for those concepts. The data values themselves will have constraints on the types of values allowed, such as strings, integers, boolean, enumerated values or controlled vocabulary terms. In addition, the FAIR Data Point specification provides an example for a standards-based approach to describe resources conform the FAIR principles, i.e. to enable uniform machine processing of centralised and federated resources[6].

[1] https://www.orpha.net/consor/cgi-bin/index.php

[2] https://rd-connect.eu/what-we-do/phenotypic-data/rb-finder-for-registries/

[3] https://samples.rd-connect.eu

[4] https://eu-rd-platform.jrc.ec.europa.eu/erdridor/

[5] https://www.iso.org/standard/31367.html

[6] https://github.com/FAIRDataTeam/FAIRDataPoint-Spec, for an example see http://purl.org/orphadata/fairpoint

The generic metadata model was further implemented to include metadata elements pertaining to the above-mentioned resources. For this purpose, an ontological model was built by using pre-existing ontologies as needed.

3. Metadata model

Initially, four key concepts that are common across all catalogs were identified: Catalogues of registries/of biobanks, Registries/Biobanks, Organizations, and Locations. Each of these concepts has sets of attributes that are shared among all instances; for example, Catalogues have titles and homepages. These are detailed in the four tables below.

In the context of FAIR, it is necessary for all concepts, and all attributes, to be identified by a globally unique and resolvable identifier. To achieve this a Uniform Resource Locator (URL[7]) was assigned to each concept/attribute, and where possible, reusing existing URLs that point to concepts/attributes in widely used Ontologies or vocabularies (for example, the "title" property of a Catalog will be indicated by the URL "http://purl.org/dc/terms/title", which is derived from the widely used Dublin Core Metadata Standard). For concepts identified as being specific to the EJP RD project, an ontology was built *de novo* (which is currently in-development and unpublished, but will be served from: http://purl.org/ejp-rd/vocabulary).

More details about these concepts, and their associated URLs and intended usage, are provided in the technical tables below.

[7] https://www.w3.org/TR/uri-clarification/

Concept name: Catalog (of rare disease registries)

Concept Unique Resource Identifier: http://purl.org/ejp-rd/vocabulary/EJPCatalogOfRegistries

Supertype Unique Resource Identifier: https://www.w3.org/ns/dcat#Catalog

Concept Description: An EJP RD Catalog is a collection of metadata about rare disease datasets

Attributes:

Attribute name	Unique Resource Identifier (URI)	Description	Range
id	http://purl.org/dc/terms/identifier	A unique identifier for the catalog	URI or String
description	http://purl.org/dc/terms/description	A textual description for the catalog	String
title	http://purl.org/dc/terms/title	The short name or title for the catalog	String
homepage	http://xmlns.com/foaf/0.1/homepage	The primary URL for the homepage of the catalog	String
publisher	http://purl.org/dc/terms/publisher	Organisation details for the catalog	Organisation
dataset	http://www.w3.org/ns/dcat#dataset	List of dataset urls	URL

Concept name: Registry dataset

Concept Unique Resource Identifier: http://purl.org/ejp-rd/vocabulary/PatientRegistryDataset

Supertype Unique Resource Identifier: https://www.w3.org/ns/dcat#Dataset

Concept Description: A dataset from a patient registry

Attributes:

Attribute name	Unique Resource Identifier	Description	Range
id	http://purl.org/dc/terms/identifier	A unique identifier for the registry	URI or String
description	http://purl.org/dc/terms/description	A textual description for the registry	String
title	http://purl.org/dc/terms/title	The short name or title for the registry	String
homepage	http://xmlns.com/foaf/0.1/homepage	The primary URL for the homepage of the dataset	String
publisher	http://purl.org/dc/terms/publisher	The organisation that published the dataset	Organisation
theme	http://www.w3.org/ns/dcat#theme	List of disease codes that describe the primary theme for the datasets	URI or String. Should be ORPHA codes.

Concept name: Organisation

Concept Unique Resource Identifier: http://xmlns.com/foaf/0.1/organisation

Concept Description: Data about an organisation responsible for publishing or producing rare disease data

Attributes:

Attribute name	Unique Resource Identifier	Description	Range
title	http://purl.org/dc/terms/title	The short name or title for the organisation	String
homepage	http://xmlns.com/foaf/0.1/homepage	The primary URL for the homepage of the organisation	String
location	http://purl.org/dc/terms/spatial	Location information for the organisation	Location

Concept name: Location

Concept Unique Resource Identifier: http://purl.org/dc/terms/LocationPeriodOrJurisdiction

Concept Description: A location, period of time, or jurisdiction.

Attributes:

Attribute name	Unique Resource Identifier	Description	Range
Country name	http://www.w3.org/2006/vcard/ns#country-name	The short name of the country	String
locality	http://www.w3.org/2006/vcard/ns#locality	The locality (e.g. city or town) associated with the address of the object	String

4. Metadata schema

In addition to describing the concepts we constructed a schema to define how instantiations of these concepts can be connected together through semantic relationships. The schema is fully specified as an OWL ontology (see later) and is also depicted in the image below (See Figure 1). Unique identifiers were assigned to each concept and relationships to ground these in standard vocabularies. The vocabulary name for each concept is prefixed to each concept name in the diagram. The catalog model is based on the existing W3C Data Catalog vocabulary (DCAT) model (https://www.w3.org/TR/vocab-dcat-2/) and is being extended for EJP needs. DCAT, in combination with the Re3Data schema (https://www.re3data.org/schema), is also the basis for the FAIR Data Point specification that is being developed to enable data analysis and machine learning across federated data resources.

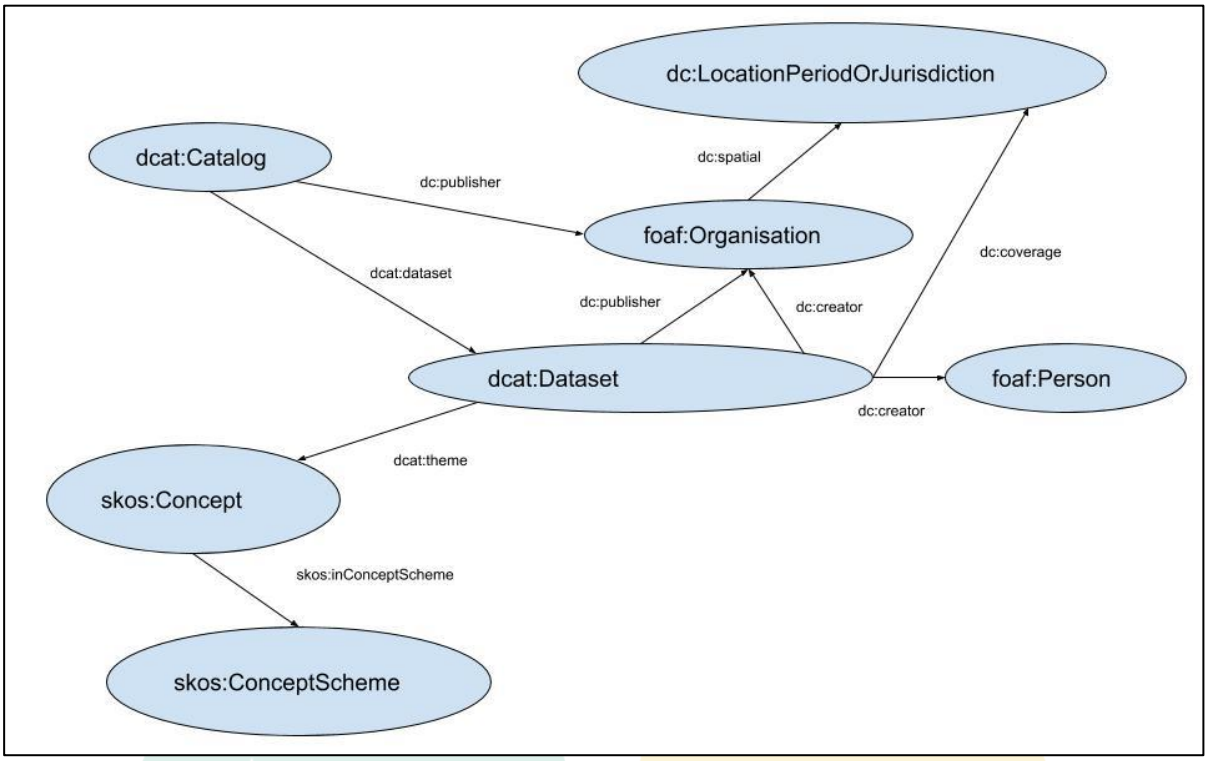

Figure 1. Representation of the major concepts in the metadata model and the relationships between them. The names are prefixed with the name of the vocabulary that defines the concept e.g. "dcat" is a reference to the Data Catalog vocabulary (https://www.w3.org/TR/vocab-dcat-2/).

The model is intended to be extensible by individual registries/resources, so more specific concepts can be introduced that extend from the base concepts. This will give the EJP RD Virtual Platform a base framework for interpreting data coming from multiple heterogeneous sources. In the next example we show how the model can be instantiated with real data from an entry in the ERDRI directory of registries. In the example we capture the ERDRI.dor (Catalog) that is published by ERDRI (Organisation), which is based in Germany (Location). Within the ERDRI.dor catalog there is an entry for the PCD Registry (Dataset), that is published by the WWU (Organisation), also based in Germany (Location). This registry has a number of themes (Disease concepts) that are described using Orphanet disease codes from the Orphanet vocabulary (Concept Scheme). See Figure 2.

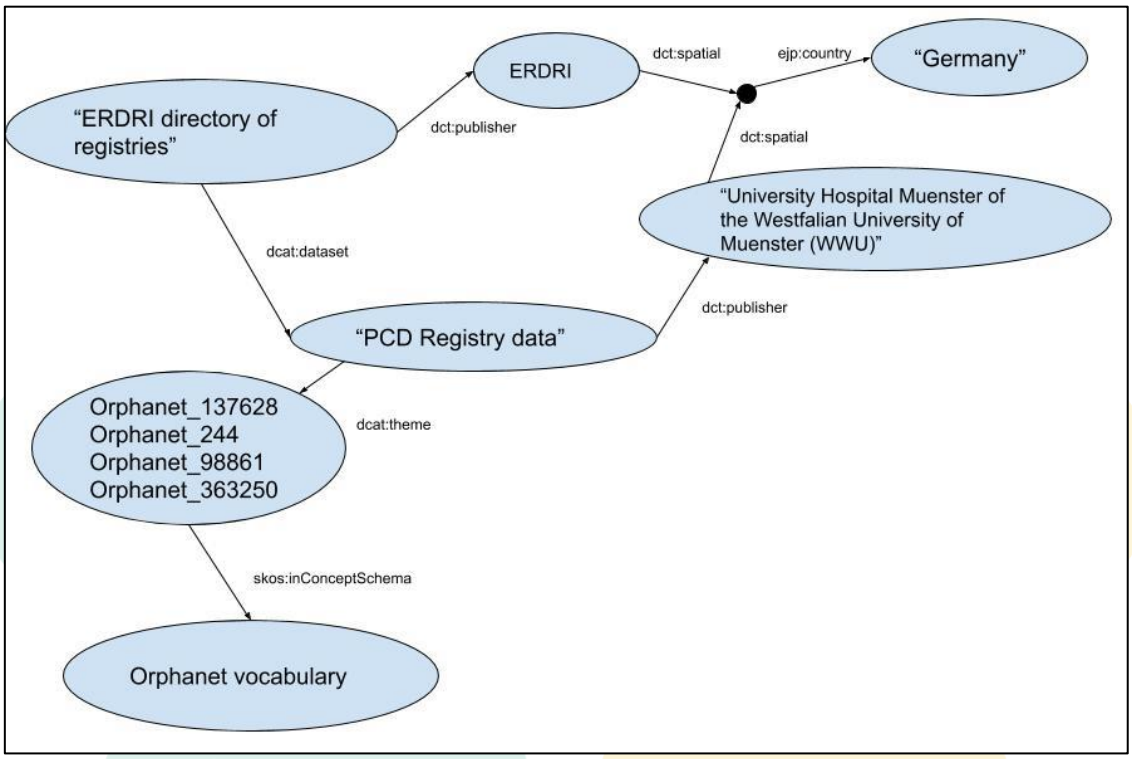

Figure 2. Representation of how the model can be instantiated with real data for the PCD registry as captured inside the ERDRI directory of registries.

Some examples of the searchable metadata are provided in Table 1.

Table 1. Examples of queryable metadata currently included in the model.

Metadata queries enabled by the model	**Example query answers**
In which Catalog is the registry/biobank's information stored?	Orphanet, RD-Connect, etc.
What is the Location (country) of the registry/biobank?	France, Italy, etc.
What Disease(s) are linked to the registry/biobank?	Rett syndrome, Usher syndrome, etc.
What is the Registry/biobank name?	REEG: Spanish Gaucher's disease registry, etc.
What is the Registry/biobank URL?	http://www.feeteg.org/G_registro.php, etc.
What is the Geographical coverage of patients included in the registry/biobank?	National, European, Global, etc.

5. Implementing the model

So far, the model has been described in abstract terms. In the spirit of FAIR, we seek to publish this model as an open standard. The core semantics of the model can be encoded in an OWL ontology. We have constructed the EJP RD ontological model that is an application ontology that pulls together concepts from mostly pre-existing ontologies to provide the overarching semantic representation of the model.

The ontological model can be used to model each concept as an OWL class and define constraints on class memberships based on properties. In addition to the concepts already presented in the core EJP RD metadata model for catalogs and patient registry datasets, the ontology includes a number of additional concepts that were found to be used in existing data registries.

We have extracted all metadata elements (or properties) registered in the ERDRI metadata repository[8] and mapped these to existing ontology concepts. These concepts have been imported via a standard pipeline to create the initial version of the EJP RD ontological model.

The EJP RD Ontological model combines imports of some biological ontologies, such as NCIT, OMIT, OMIABIS, EFO, ORDO and others. It is built to support the semantic annotations and to enhance or facilitate data retrieval in the EJP RD catalogs. The model can be viewed in any standard OWL ontology editor, such as Protege. The ontology is being developed and made available via GitHub at https://github.com/EBISPOT/EJP-Ontology.

5.1. xrJSON schema

To maximise uptake of the model we are providing a number of reference implementation that show how data can be represented in a concrete syntax for interpretation by an increased number of programs and programmers, including the EJP RD Virtual Platform. While the model is heavily grounded in Semantic Web technology standards such as RDF and OWL, it is important that simpler representations are also available, and this is consistent with the development practices of GA4GH, of which EJP is a Driver Project. We note that it is possible that mappings between existing standards to the EJP RD model will also be required as different registries will have their own data formats that will need to be handled by the EJP RD VP. We expect this to be a focus of future work.

We have initially provided a simplified view of the model based on JSON schema. JSON is an open-standard file format that is popular for transmitting data objects on the Web. JSON schema provides a vocabulary for specifying the structure of JSON documents. We can express the EJP RD catalog model in JSON schema, to provide recommended document structure for sharing data objects about patient registries and catalogs.

[8] https://eu-rd-platform.jrc.ec.europa.eu/erdri-description#inline-nav-2

Using the same example above we can express the information about the PDC registry in JSON according to our schema as follows.


```
{
 "@id": "https://www.pcdregistry.eu/",
 "@type": "PatientRegistryDataset",
 "name": "PCD Registry",
 "disease_cases": [
  {
   "disease_code": [
    {"@id": "http://www.orpha.net/ORDO/Orphanet_137628"},
    {"@id": "http://www.orpha.net/ORDO/Orphanet_363250"},
    {"@id": "http://www.orpha.net/ORDO/Orphanet_275742"},
    {"@id": "http://www.orpha.net/ORDO/Orphanet_244"},
    {"@id": "http://www.orpha.net/ORDO/Orphanet_98861"}
   ]
  }
 ],
 "publisher": [
  {
   "name": "University Hospital Muenster of the Westfalian University of Muenster (WWU)",
   "location": {
    "city": "D-48149 Muenster",
    "country": "Germany"
   }
  }
 ]
}
```


The JSON schema version of the EJP RD model is documented at https://ejp-rd-vp.github.io/resource-metadata-schema/ and the source for the schema is being developed openly on GitHub at https://github.com/ejp-rd-vp/resource-metadata-schema.

6. Future work

Deliverable 11.6 (Virtual Platform of RD resources annotated with EJP RD ontological model) provides details on how the ontological model from this deliverable has been used to develop a first prototype of the EJP RD Virtual Platform. Through developing this prototype, we identified a number of challenges when aligning catalog data to the model that will require some curatorial activity within the various catalogs to ensure the data they provide to the VP is harmonised. As the use-cases and requirements for the VP mature, there is an expectation that the EJP RD model presented here will need to be nimble to changes. In addition to the internal needs of EJP RD we are also working to ensure the model proposed by the EJP RD is interoperable with those from other stakeholder projects such as GA4GH. Similarly, we will investigate alignment with models used in other FAIR projects (e.g. IMI2, EOSC, GO FAIR implementation networks), and models that enable analysis across federated FAIR resources, such as the FAIR data point. Using common models will facilitate smooth communication between centralised and federated resources. Implementation of the models guided by data stewards will be a topic for the FAIRification task (12.3.3). As we begin to disseminate the model to a wider community of users, we will work on improving the process for community engagement and review.

To address the challenges of evolving requirements we will adopt an agile development methodology for the model. This will involve working closely with the stakeholders to collect feedback, use-cases and requirements for the VP, that will get translated into specifications for the ontological model. Regular iteration cycles coupled with early prototype development (as demonstrated in D 11.6) will be required to show the value of the proposed model and encourage adoption by the wider RD community.