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 11.2
Second Ontological model of resources metadata

Organisation name of lead beneficiary for this deliverable:
Partner 76 – ELIXIR-EMBL-EBI
Contributors: BBMRI-ERIC, GUF, INSERM-Orphanet, LUMC, UMCG, UPM

Due date of deliverable: month 24

Dissemination level:
Public

Table of Contents

Executive Summary

This deliverable describes the second ontological model of resources metadata, needed to represent other types of resources than those described in the first ontological model. Indeed, the latter focused on registry and biobank catalogues, i.e. Orphanet, RD-Connect, ERDRI.dor and BBMRI. Here, the model is extended to cover resources containing documents, services and tools. To make such resources discoverable on the EJP RD Virtual Platform, the resources' metadata need to be described. The inclusion of these new resources resulted in the need to update the first ontological model. This revision updates the first ontological model (1) by updating the base model from the Data Catalogue Vocabulary (DCAT) version 1 to 2; (2) by adding support for identifiers, quality information and citations of datasets and (3) by adding support for other types of resources, for example software. In the updated model, DCAT2 became the base standard; classes, properties and guidance are provided to address identifiers, dataset quality information, and data citation properties (in the case of research papers) and other properties to address software and research papers. This deliverable reports the detailed updates and its corresponding repository on the EJP RD GitHub repository.

1. Project Objectives

This deliverable contributes directly to the following Work Package 11 (WP11) objectives.

- It defines the second version of the resource metadata ontological model as described in WP11, Deliverable11.1.

- It updates the first ontological model "resources metadata model" to include all other resources such as software, services and documents, including research papers.
- It updates the semantic metadata model implementation integration through the EJP RD Virtual Platform.

2. Introduction

This deliverable is the result of the ongoing work initiated in Year 1, by Task 11.1 and Metadata Work Focus (WF) partners, i.e., the definition of the metadata and ontological model for describing the minimum metadata that is required to describe Rare Diseases (RD) resources (e.g., catalogues, bio-tools, research papers and documents, software) for automated rare diseases research.

During Year 1, a metadata model was developed by collecting metadata elements from existing catalogues of patient registries and biobanks. The selected catalogues for this task were the Orphanet catalogue of registries and biobanks, the RD-connect Biobank and registry finder, the RD-Connect Sample catalogue and the ERDRI directory of registries. A first version of the model was produced, based on the W3C-recommended Data Catalogue Vocabulary (DCAT) v1, and a first implementation applied to catalogues data was released on the Linked Data Platform.

This deliverable describes the year two process and release of the updated version of the EJP RD Resource Metadata Model both at the semantic level and at implementation level. The EJP RD Resources Dataset Metadata (EReMM) Model lies at the heart of the EJP RD Virtual Platform. It provides RD data sources with the information required to make resources for RD research discoverable and more **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (FAIR) in fully or partly automated scenarios, as well as to enable curators to curate their data efficiently and effectively.

3. Semantic data model

During Year 2, an updated version of the EJP RD Resources Metadata Model was produced. Following the release of the new version of DCATv2 standard in February 2020, the existing model was thoroughly reshaped accordingly.

The second version of the resources metadata model complies with the DCAT2 specification, as well as to ISO specifications, and integrates frequently used vocabularies by curators in the domain of rare diseases to the EJP RD Ontology.

The EJP RD ontology imports vocabularies from ontologies such as Experimental Factor Ontology (EFO), Semantic science Integrated Ontology (SIO), National Cancer Institute Thesaurus (NCIT[1], EDAM[2], Orphanet Rare Disease Ontology (ORDO[3], Informed

[1] https://ncithesaurus.nci.nih.gov/ncitbrowser/
[2] http://edamontology.org/page
[3] http://www.orphadata.org/cgi-bin/index.php#ontologies

Consent Ontology (ICO[4], and Information Artifact Ontology (IAO[5]The EJP RD Ontology can then be utilised at source to expose information about a RD resource in terms of a shared EJP RD vocabulary, for instance by deploying a semantic database, an RDF 'triple store', that stores specific information about the resource annotated with the model. Providing a triple store with the model as part of the EJP RD Virtual Platform (VP) offers a comprehensive and extensible vocabulary service that facilitates interoperability among rare disease

3.1. Resource Metadata Model Overview

To aid completion of a metadata entry for a new resource, and subsequently navigate content associated with an existing resource, the resource metadata model was divided into six thematic sections.

- Dataset-level information describes the properties and characteristics of the overall resource dataset – its name, a description of its content, primary language, the theme and/or subjects it encapsulates.
- Provenance information defines the origins of the resource, which is particularly useful when a resource has been provided via a third-party source.
- Legal information pertains to potential licensing arrangements, access and usage rights, etc. This will be addressed in the next update.
- Temporal information defines the time range that the dataset encapsulates.
- Spatial information describes the geospatial extent of the dataset.
- File-level information details properties of each file hosted within the dataset – its name, format (derived from file extension), size, any international standard it conforms to, etc.

3.2. The EJP RD Resources Metadata Update Scope

The core modules of the resources metadata model represent the minimum required metadata to describe resources for discoverability and interoperability on the EJP RD Virtual Platform using the metadata structure. The thematic sections are detailed in the core modules consisting of the following classes:

1. Resource[6]: this update enables the definition of a catalogue of any kind of resource. This addresses the limitation of the first version that only made provision for patient-registries and biobanks.
2. Data Service[7]: this allows for collection of operations accessible through an interface that provides access to one or more datasets or data processing functions.
3. Agent[8]: it refers to a person or organisation that is responsible for the curation and/or management of the datasets.

[4] http://www.obofoundry.org/ontology/ico.html
[5] http://www.obofoundry.org/ontology/iao.html
[6] https://github.com/ejp-rd-vp/EJP RD-vp_metadata-schemas_implementation/blob/main/doc/catalogedResources.md
[7] https://github.com/ejp-rd-vp/EJP RD-vp_metadata-schemas_implementation/blob/main/doc/dataService.md
[8] https://github.com/ejp-rd-vp/EJP RD-vp_metadata-schemas_implementation/blob/main/doc/agents.md

4. Dataset[9]: a dataset is a collection of data, published or curated by a single agent. Various data forms, including numbers, words, pixels, imagery, sound and other multi-media, and potentially other types, can be collected into a dataset.
5. Catalogue[10]: it is a kind of dataset whose member items are descriptions of datasets and data services.
6. Distribution[11]: it represents a way to access a dataset, for example, a dataset that is downloadable as a file.
7. Catalogues Record[12]: it represents a metadata item in the catalogue, primarily concerning the registration information, such as who added the item and when.

First version model

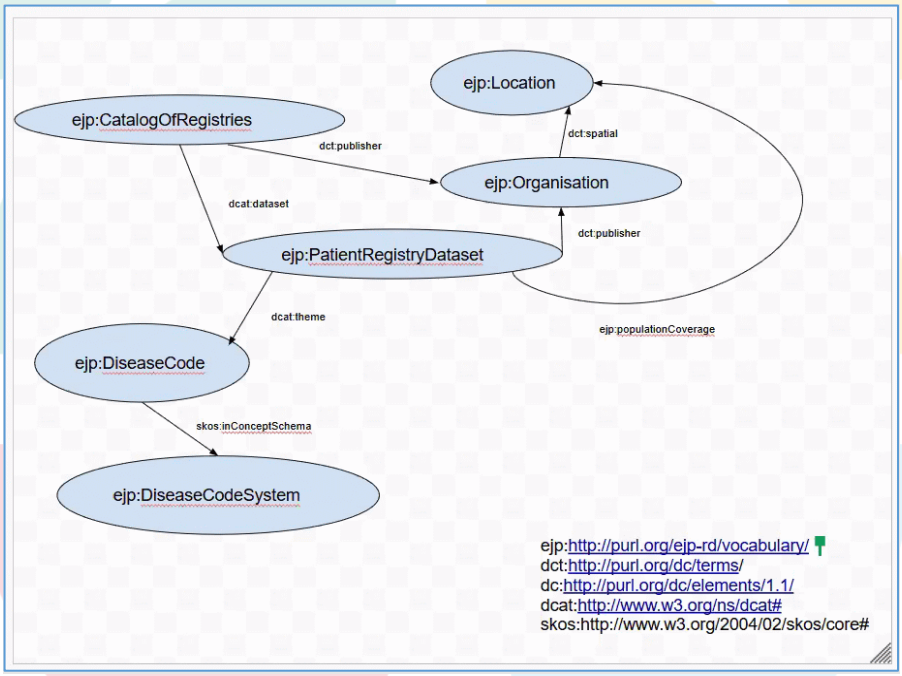

Figure 1. Resource metadata model first version.

[9] https://github.com/ejp-rd-vp/EJP_RD-vp_metadata-schemas_implementation/blob/main/doc/dataset.md

[10] https://github.com/ejp-rd-vp/EJP_RD-vp_metadata-schemas_implementation/blob/main/doc/catalog.md

[11] https://github.com/ejp-rd-vp/EJP_RD-vp_metadata-schemas_implementation/blob/main/doc/distribution.md

[12] https://github.com/ejp-rd-vp/EJP_RD-vp_metadata-schemas_implementation/blob/main/doc/agents.md

Second version model

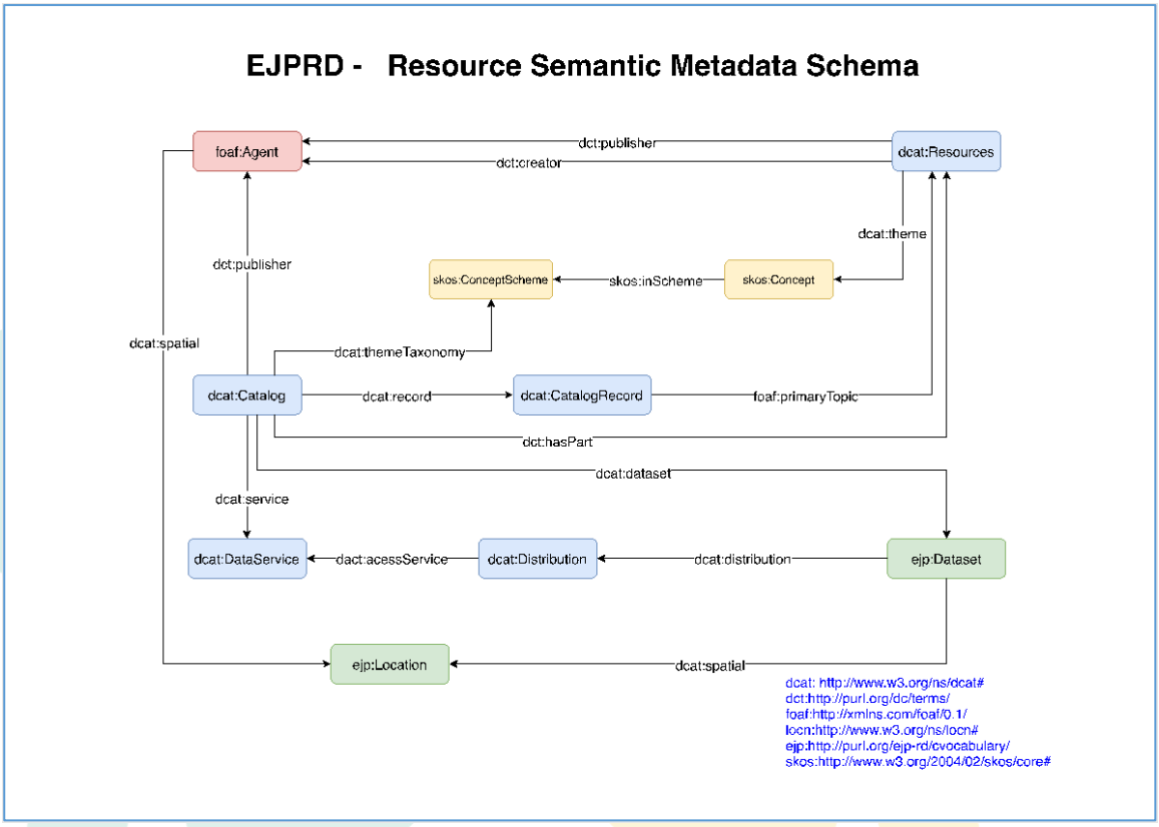

Figure 2. Resource metadata model second version.

The properties of the classes are described in detail in the GitHub repository[13]. The comprehensive framework for resources dataset metadata includes both required and optional fields. The classes properties are based primarily on the DCAT2 schema and associated namespaces with further semantics from the EJP RD-Ontology. This metadata is intended to be indexed and harvested by the EJP RD-VP search engine while resource data providers can use it to check their data compliance with the resources metadata model. The released artefact is in the release section on the repository[14].

3.3. Additional Updates to First Ontological Model

In order to make our model compliant with DCATv2, the following changes were made:

- the ejp:*CatalogofRegistries* class was replaced with the *dcat:Resource* class to refer to all resources accessible via the EJP RD virtual platform;
- the *ejp:PatientRegistryDataset* class was replaced by the *dcat:Dataset* class, which is the subclass of *dcat:Resource* class;

[13] https://github.com/ejp-rd-vp/EJP_RD-vp_metadata-schemas_implementation/blob/main/doc/homepage.md

[14] https://github.com/ejp-rd-vp/EJP_RD-vp_metadata-schemas_implementation/archive/v1.0.0.zip

- *the use of ejp:Organisation was dropped in favour of dcat:Agent,* which is the parent class of *foaf:Organisation* and *foaf:Person;*
- the *ejp:DiseaseCode* and *ejp:DiseaseCodeSystem were dropped* because they are not generic enough. Instead, the following classes were introduced: *dcat:DataService, dcat:Distribution; dcat:CatalogRecord* and *dcat:Catalog* classes.

3.4. Release of the semantic model second version

The updated model was released to the dedicated GitHub repository[15], it contains the semantic data model describing the set of minimum required metadata for describing rare disease resources to be discoverable via the EJP RD Virtual Platform. The new release contains:

1. the resource metadata model in JSON format (safe json file);
2. the semantic files, serialized in RDF turtle format describing each class of the metadata model (turtles[16] file);
3. the validation files, defined in ShEx[17] format, which ensures the validity of the implemented version of the model;
4. examples for users that illustrate how the various metadata model classes need to be instantiated for their resources;
5. the final release that is in the GitHub repository.

4. Next Steps

4.1. Automation of the Resource Semantic Metadata model

The resources semantic model will be automated; this will allow resource providers to make their data FAIR through the EJP RD virtual platform, and be able to run their data in JSON format through a web application in order to check their resource metadata compliance with the EJP RD resources metadata.

[15] https://github.com/ejp-rd-vp/EJP_RD-vp_metadata-schemas_implementation/archive/v1.0.0.zip

[16] Terse RDF Triple Language (Turtle) is a syntax and file format for expressing data in the Resource Description Framework (RDF) data model.

[17] Shape expression primer version 2 (Shape Expressions (ShEx) is a language for validating and describing RDF) - https://shex.io/shex-primer/

5. Conclusion

The second ontological model has been expanded to cover resources such as catalogues of documents and services (like the EJP RD WP19 Innovation Management Toolbox, EATRIS and ECRIN catalogues) and is now being able to include further resources in the coming years.

References

EDAM ontology	http://edamontology.org/page
Experimental Factor Ontology (EFO)	https://www.ebi.ac.uk/efo/
Information Artifact Ontology (IAO)	http://www.obofoundry.org/ontology/iao.html
Informed Consent Ontology (ICO)	http://www.obofoundry.org/ontology/ico.html
JSON	https://tools.ietf.org/html/rfc4627
National Cancer Institute Thesaurus (NCIT)	https://ncithesaurus.nci.nih.gov/ncitbrowser/
Orphanet Rare Disease Ontology (ORDO)	http://www.orphadata.org/cgi-bin/index.php#ontologies
RDF	https://www.w3.org/RDF/
Semantic science Integrated Ontology (SIO)	https://www.ebi.ac.uk/ols/ontologies/sio
SHACL	*Shapes Constraint Language (SHACL)*. Holger Knublauch; Dimitris Kontokostas. W3C. 20 July 2017. W3C Recommendation. URL: https://www.w3.org/TR/shacl/
ShEx	*Shape Expressions Language 2.1*. Shape Expressions W3C Community Group. 17 November 2018. Draft Community Group Report. URL: http://shex.io/shex-semantics/

Abbreviations

DCAT - Data Catalogue Vocabulary

ECRIN - The European Clinical Research Infrastructure Network

EJP RD - European Joint Programme on Rare Diseases

JSON - JavaScript Object Notation

ERN - European Reference Networks

RDF - Resource Description Framework