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 11.3
Third 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 36

Dissemination level:
Public

Table of Contents

1. Project Objectives

This deliverable describes the outputs for WP11, Task 1, Subtask 1 & 2 for Year 3 that we made to the resource metadata model and the resource metadata ontology in support of the data resources: Cellosaurus, hPSCReg, Bio.tools and WikiPathways:

- Updates to the resource metadata model of Year 2 have been continued in Year 3 in support of Cellosaurus, hPSCReg, Bio.tools and WikiPathways.
- In accordance with the resource metadata model, instance data were created to describe Cellosaurus, hPSCReg, Bio.tools and WikiPathways.
- In Year 1 an ontological representation has been created of the resource metadata model based on our first landscape analysis. This ontology has now been updated to reflect our most up to date knowledge of the EJP RD needs.
- We have created software that can serve as an endpoint that can be used for validating data describing resource metadata against the resource metadata model for determining adherence.

These updates are described in detail below.

2. Semantic data model

In this section we describe the updates to the resource metadata model in detail. To follow this, it necessary to understand the difference between the resource metadata model expressed as ShEx shapes, the ontology representing the model, instances adhering to the ShEx shapes and the validator for validating that an instance adheres to a particular ShEx shape. As a small example, we will consider location information.

The ShEx shape is defined as follows:

```
PREFIX dct:   <http://purl.org/dc/terms/>
:locationShape IRI {
  a [dct:Location];
  dct:title xsd:string;
  dct:description xsd:string*;
}
```

This states that a location must be a dct:Location and it must have a title defined using dct:title and it may have a description defined using dct:description.

An instance that adheres to this shape may look as follows:

```
:location a dct:Location ;
  dct:title  "Fraunhofer-Institut für Biomedizinische Technik (IBMT)";
  dct:description "Anna-Louisa-Karsch Str. 2, 10178 Berlin, Germany" .
```

The validator described in Section 2.4 can validate an instance against a shape and provide a report the findings of the validation.

The ShEx shapes provide registries with the information necessary to describe their resources (as instance data) in a way that is meaningful to the EJP RD virtual platform, thereby making their resources searchable on EJP RD virtual platform. Using the validator, registries can determine whether their instance data describing their resources, adhere to the EJP RD requirements.

To enable crosswalks in future, the resource metadata ontology is included. All classes and properties used in the resource metadata model are present in the resource metadata ontology. This ontology can in time enable the possibility for crosswalks. With regards to our example, the resource metadata ontology includes the dct:Location class. We could in future add mapping information to this class to location definitions based on other ontologies.

2.1. Resource Metadata Model Updates

The resource metadata model of Year 2[1] is given in Fig. 1 and the updated version[2] is given in Fig. 2. The key semantic changes that have been made to the ShEx[3] shapes representing resource metadata model are:

- Where possible, we refined the concepts we use enabling more precise representation of concepts. Thus, instead of using the more general concept foaf:Agent, we gave preference to the use of the more specific concept foaf:Organization.
- We gave preference to using concepts from well-established vocabularies rather than introducing our own to promote interoperability of EJP RD with other resources (the I in Findable, Accessible, Interoperable and Re-usable). This resulted in the replacement of ejp:Location by dct:Location
- We started using the Semantic Science Integrated Ontology[4] is_related_to (sio:SIO_000001) property and annotation (sio:SIO_001166) to represent respectively related and annotation information that is not necessarily precise. For example, in the Orphanet catalog[5] patient registries and biobanks are annotated with population coverage which can be treated as annotations about a resource and connect to resource with a generic object property is_related_to.
- Where EJP RD has identified project specific needs, new concepts have been introduced. These concepts are:
    - ejp:PatientRegistry: This class is used to describe a patient registry. As part of capturing rare disease information EJP RD is interested in patient data. This class was introduced since no appropriate class could be found in any existing ontologies.
    - ejp:Biobank: This class is used to describe a biobank. We have decided to create our own class instead of using the Biobank class that is available in the EDAM[6] ontology. The reason for this is that we decided to base the resource metadata model on DCAT version 2.0[7] in Year 2. EDAM is not based on DCAT 2.0. Hence, the reason for creating our own ejp:Biobank class. We used DCAT here as the larger EJP RD project has decided to use DCAT in Year 2.

[1] https://github.com/ejp-rd-vp/resource-metadata-schema
[2] https://github.com/ejp-rd-vp/resource-metadata-schema/tree/henriette_develop
[3] https://shex.io/shex-semantics/
[4] https://semanticscience.org/
[5] https://www.orpha.net/consor/cgi-bin/index.php
[6] http://edamontology.org
[7] https://www.w3.org/TR/vocab-dcat-2/

- ejp:Guideline: This is a class to describe a guideline. In Year 2 IMT[8] has been added as a resource that deals with guidelines with regard to drug development.
- ejp: populationCoverage: This data property with range xsd:string describes the population that is represented by a patient registry. Current permissible values are "National", "International" and "Regional". We intent to replace this in future with an ejp:populationCoverage object property with range a class called ejp:PopulationCoverage that consist only of the individuals National, International and Regional. In this new representation National, International and Regional each will have its own IRI thereby unambiguously identifying the population coverage of a patient registry.

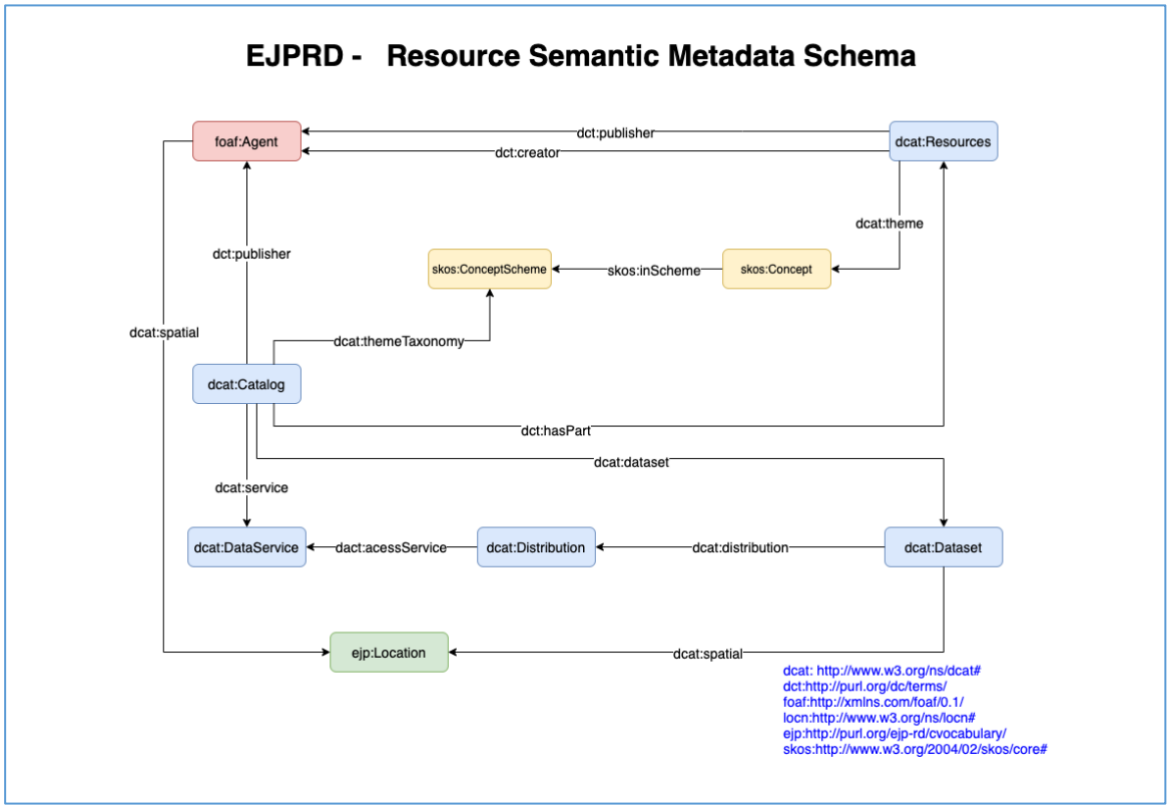

Figure 1. Year 2 resource metadata model

[8] https://ejprd.getyos.com

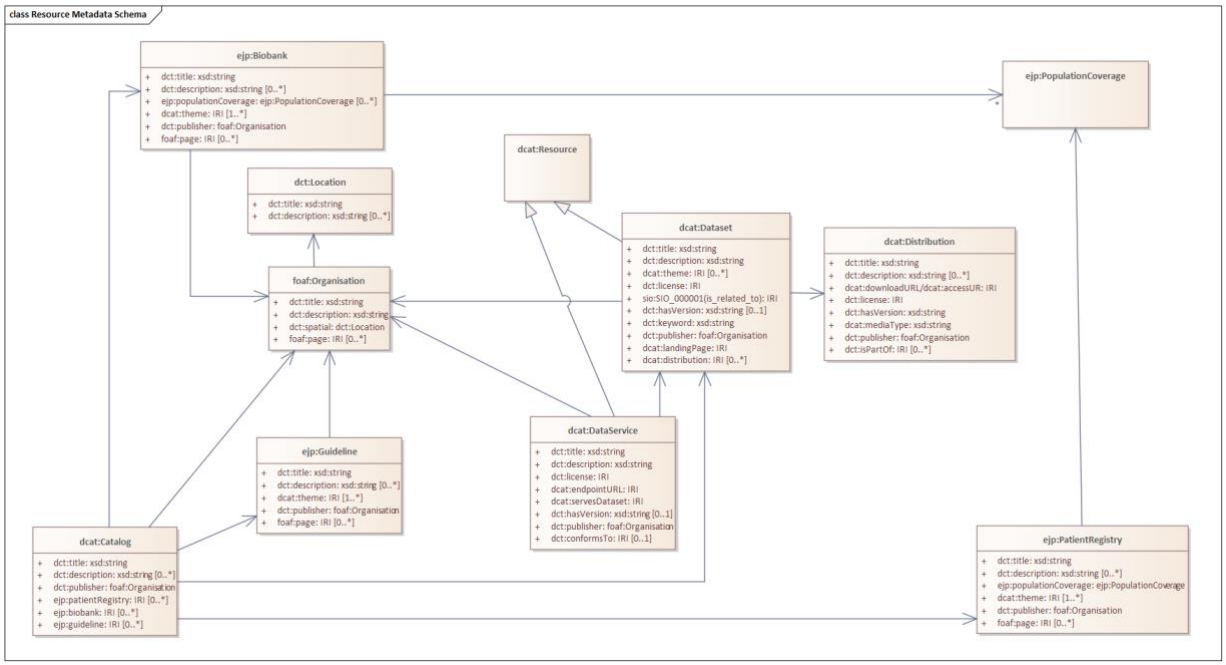

Figure 2. Year 3 resource metadata model

2.2. Resource Metadata Instances

As part of the continued efforts of Subtask 1 & 2, Task 1 of WP11, the reach of EJP RD have been extended to include access to Cellosaurus, hPSCreg, Bio.tools and WikiPathways. To this end resource metadata instances[9] have been added to describe these resources. A brief summary of dataset and data service information is given for each of these resources in Table 1.

Table 1. Dataset and data service locations for new resources

Resource	Dataset location	Dataservice location
bio.tools	https://bio.tools	https://bio.tools/api/tool/
Cellosaurus	https://web.expasy.org/cellosaurus/	https://web.expasy.org/cgi-bin/cellosaurus/search
hPSCreg	https://hpscreg.eu/export/cell_lines/	https://hpscreg.eu/api/names
WikiPathways	https://www.wikipathways.org	http://sparql.wikipathways.org/sparql

2.3. Resource Metadata Ontology Update

In Year 3 we did a complete revamp of the resource metadata ontology. The reason for this revamp stems from the fact that our approach to the design of the resource metadata ontology has changed and our understanding of the project needs has matured. In the initial development of this ontology in Year 1 the exact needs of the EJP RD have been vague and therefore the decision was made to create an ontology that has as wide a reach as is possible from a rare disease perspective, including all high-level entities of relevance to the domain. In Year 2 we continued on this path, but in this year, we reviewed our initial decision to focus on needs which have emerged as use cases become concrete. Our findings indicated:

[9] https://github.com/ejp-rd-vp/resource-metadata-schema-instances

- We model 3000 classes and many of these are not used in EJP RD. As an example, Orphanet:194 is an obsolete class that formed part of the original ontology, which we now have removed in the revamped ontology.
- The correlation between the resource metadata model and its ontology has been unclear and this confused collaborators. In particular the resource metadata model uses FOAF[10], DCAT, Dublin Core[11] and SIO while the original ontology also used for example EDAM and EFO[12]. The resource metadata ontology is now also based on only FOAF, DCAT, Dublin Core and SIO.

In Fig. 3 we compare key metrics between our initial ontology designed and the revamped design where one can see, for example, that we reduced the number of classes from 3195 to 38. This substantial reduction in classes means it will be much easier to maintain this ontology in future.

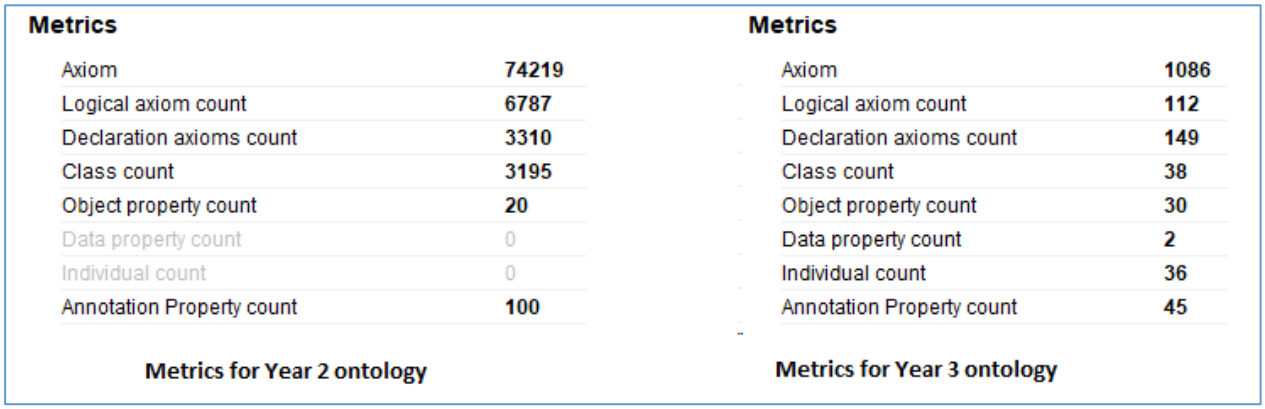

Figure 3. Metrics of initial ontology design versus revamped ontology

As part of the redesign of the ontology we decided to move the ontology out from the resource metadata model[13] GitHub repository into its own repository[14]. The reason for this is that the reasons for changing the resource metadata model is likely to be different from the reasons for changing the resource metadata ontology. I.e., the resource metadata model may need to change due to small tweaks to the ShEx shapes, which does not affect the ontology. An example of such a change is if we need to make ejp:populationCoverage mandatory rather optional. Similarly, annotations in the ontology may need to be updated which have no related representation in the ShEx shapes of the resource metadata model. An example is if we want to refine the definition of ejp:PatientRegistry. This change will have no effect on the ShEx shapes of the resource metadata model.

2.4. Resource Metadata Validator

As part of our Year 3 output, we have created software[15] that can serve as a REST endpoint to validate RDF data in Turtle syntax against a ShEx shapes file. This service can be used to

[10] http://xmlns.com/foaf/spec/
[11] https://dublincore.org/schemas/
[12] https://www.ebi.ac.uk/efo/
[13] https://github.com/ejp-rd-vp/resource-metadata-schema/tree/master/EJP-Ontology
[14] https://github.com/ejp-rd-vp/resource-metadata-schema-ontology
[15] https://github.com/ejp-rd-vp/resource-metadata-validator/

validate whether a potential EJP RD resource, described in RDF, adheres to the resource metadata model of EJP RD, described as a ShEx shape.
In this initial version of the validator emphasis has been placed on ensuring that it can validate data against all ShEx shapes that adhere to the ShEx specification, as well as, the ShEx shapes used by the resource metadata model (See Table 2). Extensive tests in this regard were merited by the fact that earlier attempts to create such a service were unable to cater for the variety of validation that are required by the resource metadata model. In particular initial attempts had difficulty in dealing with any shapes referencing other shapes.

Table 2. Resource metadata validator test locations

Test type	Location of tests
ShEx specification tests	https://github.com/ejp-rd-vp/resource-metadata-validator/tree/master/src/test/resources/shexprimer
Resource metadata model tests	https://github.com/ejp-rd-vp/resource-metadata-validator/tree/master/src/test/resources/metamodel

3. Next Steps

In Year 4 we plan to continue working with our collaborators of WP 11, Task 1, Subtask 1&2 to expand the resource metadata model, the resource metadata ontology, and the resource metadata validator to new EJP RD resources thereby enabling registries to make their resources searchable on the EJP RD platform.

4. Conclusion

We described the changes made to the resource metadata model, the resource metadata ontology in support of Cellosaurus, hPSCReg, Bbio.tools and WikiPathways and the introduction of the resource metadata validator. In summary the repository location of these changes is given in Table 3.

Table 3. Repository locations for Year 3 deliverables

Deliverable	Y2	Y3
Resource metadata model	https://github.com/ejp-rd-vp/resource-metadata-schema	https://github.com/ejp-rd-vp/resource-metadata-schema/tree/henriette_develop
Resource metadata ontology	https://github.com/ejp-rd-vp/resource-metadata-schema/tree/master/EJP-Ontology	https://github.com/ejp-rd-vp/resource-metadata-schema-ontology
Resource metadata instances	Did not exist	https://github.com/ejp-rd-vp/resource-metadata-schema-instances
Resource metadata validator	Did not exist	https://github.com/ejp-rd-vp/resource-metadata-validator/

Glossary

DCAT: The Data Catalog Vocabulary is a W3C specification for describing datasets and dataservices. See https://www.w3.org/TR/vocab-dcat-2/.

Dublin Core: It is a set of specifications for describing resources. See https://dublincore.org/.

FOAF: It is a specification for describing persons, their activities and relations to people and objects. See http://xmlns.com/foaf/spec/.

Orphanet catalog: It is a portal for rare disease and orphan drug. See https://www.orpha.net/consor/cgi-bin/index.php.

ShEx: Also referred to as the ShEx specification. The Shape Expressions (ShEx) language is a specification that describes RDF nodes and graph structures. See https://shex.io/shex-semantics/.

ShEx shape: A ShEx shape defines the expected structure of RDF data based on the ShEx specification.