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 11.5
Fifth Ontological model of resources metadata

Organisation name of lead beneficiary for this deliverable:
Partner 76 – ELIXIR/EMBL-EBI

Contributors: GUF, INSERM-Orphanet, LUMC, UPM

Due date of deliverable: month 60

Dissemination level:
Public

Table of Contents

1. Introduction

The EJP RD Virtual Platform (VP) is an ecosystem of rare disease related data resources and auxiliary services via which users can automatically find, access, integrate and use data for various purposes, ranging from discovering data and samples to analysing data. That is, the VP is a system to help making rare disease data and services Findable, Accessible, Interoperable and Reusable (FAIR).

This deliverable reports on the progress of the fifth year (Y5) of the work pertaining to subtasks 11.1.1 and 11.1.2 of Work Package 11. The objective of Work Package 11 is to develop a resource metadata model and an ontological model to support the VP. The impact of the resource metadata model and the ontological model is that it provides a common resource metadata specification which allows a resource that adheres to the specification to be onboarded onto the VP, thereby making its data and services FAIR via the VP. In this section we provide a summary of our work of Years 1-4, then we list the key updates for Year 5 and lastly, we describe the structure of the rest of this document.

In Year 1 we provided initial resource metadata- and ontological models based on concepts that are common to all rare disease resources: Catalogues (of registries/of biobanks), Registries/Biobanks, Organizations, and Locations. For the definition of Catalogues we referred to DCAT version 1. In Year 2 we recognised that even though these concepts are common across rare disease resources, there are substantial richness and differences between resources that our model did not cater for. This prompted us to adopt DCAT version 1 as a basis for our resource model. This resulted in the inclusion of concepts like Dataset, Data Service and Distribution from DCAT version 1. In Year 3 we updated our model to correspond with DCAT version 2 and extended it based on our initial efforts in onboarding resources like WikiPathways, bio.tools, Cellosaurus and hPSCreg. In Year 4 we distinguished between datasets that are discoverable and datasets that are both discoverable and queryable, we aligned the resource metadata model and the resource metadata ontology closely with the FAIR Data Point (FDP) specification and we provided a spreadsheet and related documentation for the manual onboarding of resources to the VP. Distinguishing between datasets that are only discoverable versus datasets that are both discoverable and queryable enable to the VP provide the appropriate level of access to resources that only want to be discoverable by the VP and not queryable as well, whilst alignment with the FDP specification enable resources to be onboarded to the VP through making their data available as an FDP. The spreadsheet is for those resources who prefer to be manually onboarded rather than using a technological solution like FDP.

In Year 5 our changes to the resource metadata model and the resource metadata ontology mainly related to the changing understanding of the requirements of resources to be onboarded (i.e., the need of organisations and datasets to have logo associated with their organisation or dataset) and alignment with the DCAT version 3. What has become clear is the need for governance of changes made to the resource metadata model and the resource metadata ontology to ensure that the implications for consumers of these artefacts are taken into consideration. Even minimal changes in the metadata model performed this year have resulted in cascaded impacts on the services that consume the model, including FAIR Data Point implementations and backend querying facilities of the Virtual Platform Portal that use beacon2-compliant API services. For example, when foaf:page was changed to dcat:landingPage, the VP, FDP and manual resource upload all needed to implement this change, in order to provide a seamless experience across these services.

The rest of this report is structured as follows: In the next Section 2 we discuss the approach we followed in creating the outputs for this deliverable. Section 3 discusses in detail the updates to the resource metadata model and the related ontological model as well as the governance process used for requesting and making changes to the model and the ontology.

2. Resource Metadata Updates

This section details the changes of year 5 to the resource metadata model and the resource metadata ontology. These changes have also been applied to the associated for onboarding resources manually.

2.1 Approach

The work of Year 5 builds on the work that has been done in years 1-4. Updates to the resource metadata model and ontological model have been informed by workshops, meetings with resources, collaboration with EJP RD colleagues and various EJP RD project meetings. Moreover, much of the input has come from the efforts of EJP RD teams in onboarding resources. Then there are some of the changes that stems from the changes from DCAT version 2 to version 3.

.

2.2 Resource metadata model updates

In this section we list and describe the changes made to the resource metadata model. For reference Figure 1 shows the resource metadata model as it was at the end of year 4 and in Figure 2 the model is given as it is at the end of year 5.

The reasons for changes to the resource metadata model are the following:

1. An existing property or class needed to change to reflect the latest requirements of the VP.
2. Additional properties and / or classes needed to be added based on the needs of the VP.
3. Changes were made due to migrating from DCAT version 2 to DCAT version 3.

These modifications are described in the following sections.

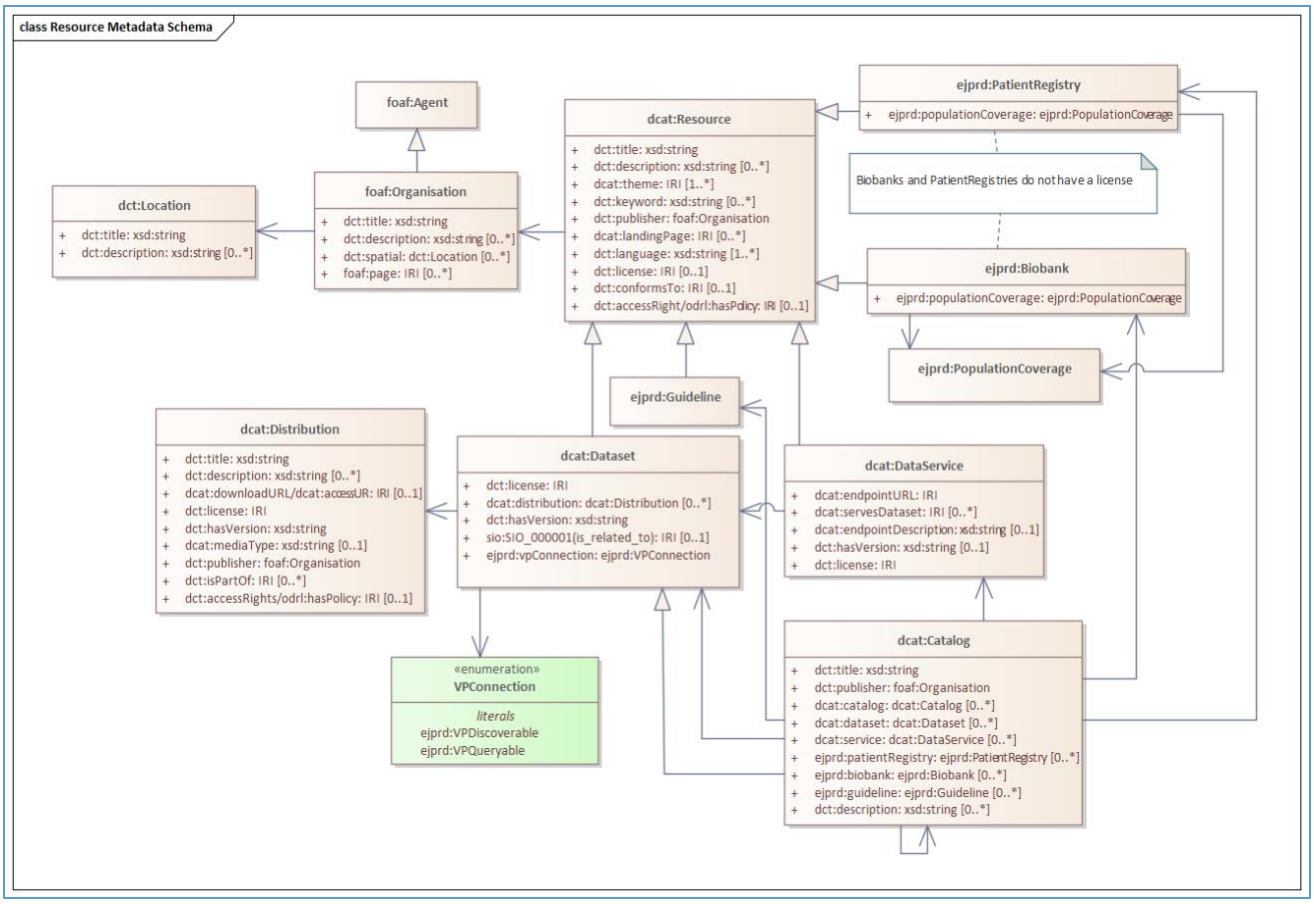
class Resource Metadata Schema
foaf:Agent
dct:Location
+ dct:title: xsd:string
+ dct:description: xsd:string [0..*]
foaf:Organisation
+ dct:title: xsd:string
+ dct:description: xsd:st ring [0..*]
+ dct:spatial: dct:Location [0..*]
+ foaf:page: IRI [0..*]
dcat:Resource
+ dct:title: xsd:string
+ dct:description: xsd:string [0..*]
+ dcat:theme: IRI [1..*]
+ dct:keyword: xsd:string [0..*]
+ dct:publisher: foaf:Organisation
+ dcat:landingPage: IRI [0..*]
+ dct:language: xsd:string [1..*]
+ dct:license: IRI [0..1]
+ dct:conformsTo: IRI [0..1]
+ dct:accessRight/odrl:hasPolicy: IRI [0..1]
ejprd:PatientRegistry
+ ejprd:populationCoverage: ejprd:PopulationCoverage
Biobanks and PatientRegistries do not have a license
ejprd:Biobank
+ ejprd:populationCoverage: ejprd:PopulationCoverage
ejprd:PopulationCoverage
ejprd:Guideline
dcat:Distribution
+ dct:title: xsd:string
+ dct:description: xsd:string [0..*]
+ dcat:downloadURL/dcat:accessUR: IRI [0..1]
+ dct:license: IRI
+ dct:hasVersion: xsd:string
+ dcat:mediaType: xsd:string [0..1]
+ dct:publisher: foaf:Organisation
+ dct:isPartOf: IRI [0..*]
+ dct:accessRights/odrl:hasPolicy: IRI [0..1]
dcat:Dataset
+ dct:license: IRI
+ dcat:distribution: dcat:Distribution [0..*]
+ dct:hasVersion: xsd:string
+ sio:SIO_000001(is_related_to): IRI [0..1]
+ ejprd:vpConnection: ejprd:VPConnection
dcat:DataService
+ dcat:endpointURL: IRI
+ dcat:servesDataset: IRI [0..*]
+ dcat:endpointDescription: xsd:string [0..1]
+ dct:hasVersion: xsd:string [0..1]
+ dct:license: IRI
«enumeration»
VPConnection
literals
ejprd:VPDiscoverable
ejprd:VPQueryable
dcat:Catalog
+ dct:title: xsd:string
+ dct:publisher: foaf:Organisation
+ dcat:catalog: dcat:Catalog [0..*]
+ dcat:dataset: dcat:Dataset [0..*]
+ dcat:service: dcat:DataService [0..*]
+ ejprd:patientRegistry: ejprd:PatientRegistry [0..*]
+ ejprd:biobank: ejprd:Biobank [0..*]
+ ejprd:guideline: ejprd:Guideline [0..*]
+ dct:description: xsd:string [0..*]

Figure 1: Year 4 resource metadata model

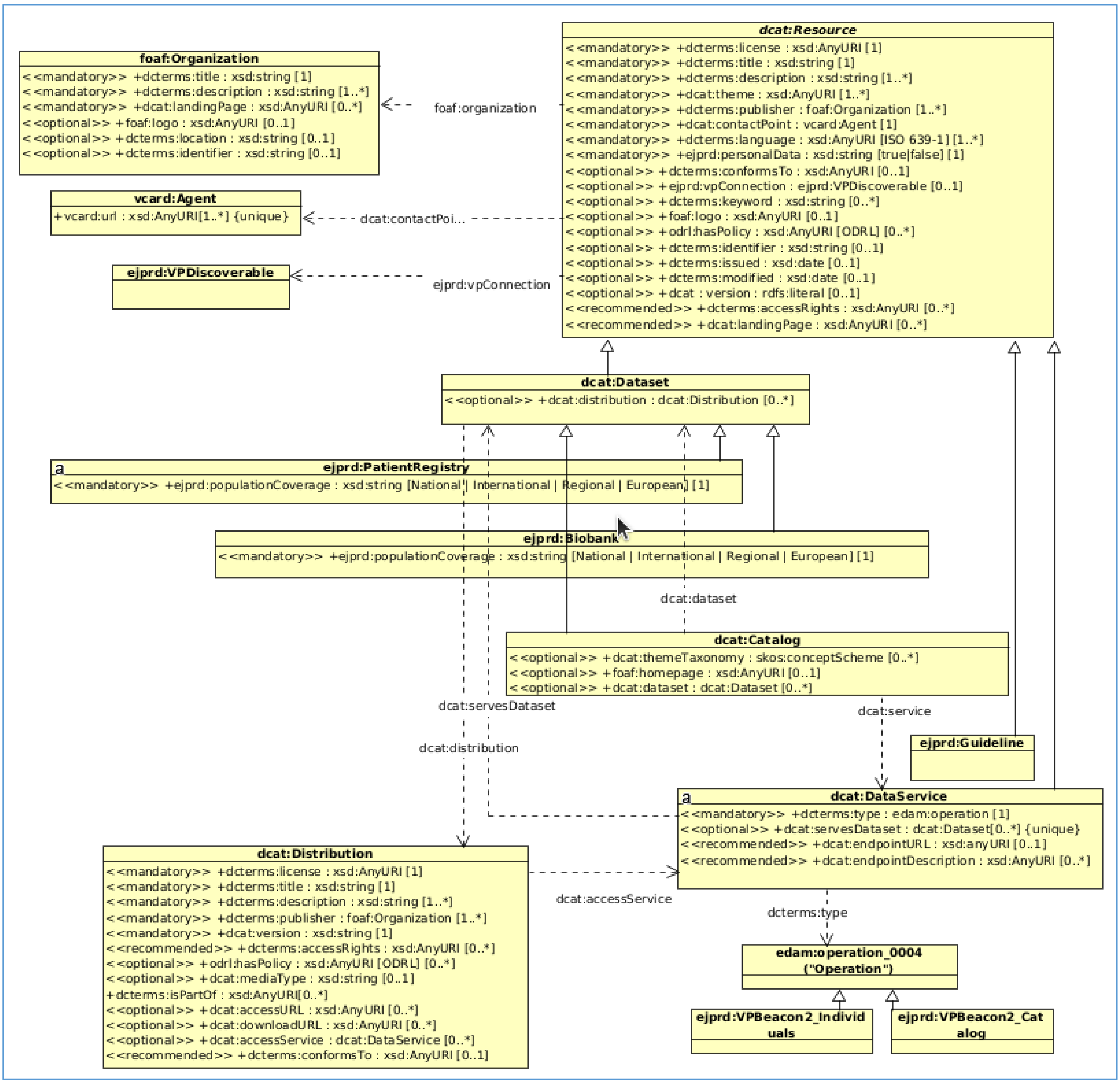
dcat:Resource
<<mandatory>> +dcterms:license : xsd:AnyURI [1]
<<mandatory>> +dcterms:title : xsd:string [1]
<<mandatory>> +dcterms:description : xsd:string [1..*]
<<mandatory>> +dcat:theme : xsd:AnyURI [1..*]
<<mandatory>> +dcterms:publisher : foaf:Organization [1..*]
<<mandatory>> +dcat:contactPoint : vcard:Agent [1]
<<mandatory>> +dcterms:language : xsd:AnyURI [ISO 639-1] [1..*]
<<mandatory>> +ejprd:personalData : xsd:string [true|false] [1]
<<optional>> +dcterms:conformsTo : xsd:AnyURI [0..1]
<<optional>> +ejprd:vpConnection : ejprd:VPDiscoverable [0..1]
<<optional>> +dcterms:keyword : xsd:string [0..*]
<<optional>> +foaf:logo : xsd:AnyURI [0..1]
<<optional>> +odrl:hasPolicy : xsd:AnyURI [ODRL] [0..*]
<<optional>> +dcterms:identifier : xsd:string [0..1]
<<optional>> +dcterms:issued : xsd:date [0..1]
<<optional>> +dcterms:modified : xsd:date [0..1]
<<optional>> +dcat : version : rdfs:literal [0..1]
<<recommended>> +dcterms:accessRights : xsd:AnyURI [0..*]
<<recommended>> +dcat:landingPage : xsd:AnyURI [0..*]
foaf:Organization
<<mandatory>> +dcterms:title : xsd:string [1]
<<mandatory>> +dcterms:description : xsd:string [1..*]
<<mandatory>> +dcat:landingPage : xsd:AnyURI [0..*]
<<optional>> +foaf:logo : xsd:AnyURI [0..1]
<<optional>> +dcterms:location : xsd:string [0..1]
<<optional>> +dcterms:identifier : xsd:string [0..1]
foaf:organization
vcard:Agent
+vcard:url : xsd:AnyURI[1..*] {unique}
dcat:contactPoi...
ejprd:VPDiscoverable
ejprd:vpConnection
dcat:Dataset
<<optional>> +dcat:distribution : dcat:Distribution [0..*]
a
ejprd:PatientRegistry
<<mandatory>> +ejprd:populationCoverage : xsd:string [National | International | Regional | European] [1]
ejprd:Biobank
<<mandatory>> +ejprd:populationCoverage : xsd:string [National | International | Regional | European] [1]
dcat:dataset
dcat:Catalog
<<optional>> +dcat:themeTaxonomy : skos:conceptScheme [0..*]
<<optional>> +foaf:homepage : xsd:AnyURI [0..1]
<<optional>> +dcat:dataset : dcat:Dataset [0..*]
dcat:servesDataset
dcat:distribution
dcat:service
ejprd:Guideline
a
dcat:DataService
<<mandatory>> +dcterms:type : edam:operation [1]
<<optional>> +dcat:servesDataset : dcat:Dataset[0..*] {unique}
<<recommended>> +dcat:endpointURL : xsd:anyURI [0..1]
<<recommended>> +dcat:endpointDescription : xsd:AnyURI [0..*]
dcat:Distribution
<<mandatory>> +dcterms:license : xsd:AnyURI [1]
<<mandatory>> +dcterms:title : xsd:string [1]
<<mandatory>> +dcterms:description : xsd:string [1..*]
<<mandatory>> +dcterms:publisher : foaf:Organization [1..*]
<<mandatory>> +dcat:version : xsd:string [1]
<<recommended>> +dcterms:accessRights : xsd:AnyURI [0..*]
<<optional>> +odrl:hasPolicy : xsd:AnyURI [ODRL] [0..*]
<<optional>> +dcat:mediaType : xsd:string [0..1]
+dcterms:isPartOf : xsd:AnyURI[0..*]
<<optional>> +dcat:accessURL : xsd:AnyURI [0..*]
<<optional>> +dcat:downloadURL : xsd:AnyURI [0..*]
<<optional>> +dcat:accessService : dcat:DataService [0..*]
<<recommended>> +dcterms:conformsTo : xsd:AnyURI [0..1]
dcat:accessService
dcterms:type
edam:operation_0004
("Operation")
ejprd:VPBeacon2_Individuals
ejprd:VPBeacon2_Catalog

Figure 2. Year 5 EJP RD resource metadata model[1]

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

2.2.1 Changes to existing properties and classes

In this section we detail changes that have been made to existing properties and classes of the resource metadata model.

- All ranges of type IRI have been changed to be of the type xsd:AnyURI as IRI is not a recognised type in any namespace. This is also in accordance with the DCAT specification.
- All dcterms:description properties were made mandatory in support of a meaningful user experience.
- All dcterms:accessRights and odrl:hasPolicy have been split to make clear that these are different concerns. For both, the maximum cardinality has been changed to allow for associating an arbitrary number of documents.
    - dcterms:accessRights refers to information about who can access the resource or an indication of its security status. This should point to a URL where this information can be found. Specifying this property is recommended but not mandatory.
    - odrl:hasPolicy refers to an ODRL conformant policy document (https://www.w3.org/TR/odrl-model/) expressing the rights and/or responsibilities associated with access to and/or use of the resource.
- dcat:Resource
    - The maximum cardinality of dcterms:publisher have been changed to allow an arbitrary number of publishers of a resource.
    - It is now recommended to specify a dcat:landingPage.
- foaf:Organization
    - foaf:page was replaced by dcat:landingPage which is better aligned with the DCAT specification.
    - The range of dcterms:location was changed to xsd:string and the dependency on the dcterms:Location class was removed.
- dcat:Dataset
    - dcat:endpointURL is now recommended rather than mandatory
    - dcat:servesDataset now must be unique within the VP.
- dcat:Distribution
    - The maximum cardinality of dcterms:publisher was changed to allow for an arbitrary number of publishers per distribution.
    - dcat:accessURL and dcat:downloadURL have been split to indicate that these represent different concerns. That is, dcat:accessURL represents a URL that gives access to a distribution of the dataset, i.e. a landing page, feed or SPARQL endpoint. dcat:downloadURL represents a URL from where the data can be downloaded.

2.2.2 Properties and classes that were added

Some new classes and properties were introduced, which detailed in this section.

- dcat:Resource
    - ejprd:personalData with the range consisting of the values “true” or “false” was added. It should be set to "true" if the resource to be onboarded to the VP contains personal data, otherwise "false". Personal data in a VP context refers to data related to identified or identifiable persons (as per GDPR definition).

  - dcat:contactPoint with range vcard:Agent has been added to capture contact information. The vcard:Agent class has been added and expects that at least one vcard:url will be supplied that is unique within the VP.
  - ejprd:vpConnection with range ejprd:VPDiscoverable have been added which resources can use to tag their metadata with to indicate that this metadata will be accessible on the VP. Metadata for which this tag is not specified will not be accessible on the VP.
  - An optional foaf:logo property has been added to allow association of a graphic with the resource.
  - An optional dcterms:identifier property was added to associate a unique reference with a resource.
  - An optional dcterms:issued property with range xsd:date was added to record the resource publication date.
  - An optional dcterms:modified has been added to record the date on which the resource was last updated or changed.
- foaf:Organisation
  - An optional foaf:logo was added to associate a logo with an organisation.
  - An optional dcterms:identifier was added to provide a unique identifier to an organisation.
- dcat:Distribution
  - dcterms:conformsTo has been added to indicate an established standard to which the distribution conforms.
- dcat:DataService
  - The mandatory property dcterms:type with range edam:operation_004 (Operation) was added.
  - API specifications for two types of beacon2-compliant endpoints were designed to support discoverability of resources in the EJP RD Virtual Platform: an endpoint for safe record-level discovery and another for resource-level querying. For this purpose two classes were added, namely ejprd:VPBeacon2_Individuals and ejprd:VPBeacon2_Catalog, both as subclasses of edam:operation_004. For Beacon services these 2 classes are used while for all other services subclasses of edam:operation_004 are to be used. ejprd:VPBeacon2_Individuals indicates that the annotated service provides the Beacon2 /individuals interface and ejprd:VPBeacon2_Catalog indicates that the annotated service provides the Beacon2 /catalog interface.dcat:Catalog .
  - An optional dcat:themeTaxonomy has been added to define the taxonomy used to classify resources documented in the catalog.
  - An optional foaf:homepage has been added to define the homepage of a catalog.

2.2.3 DCAT version 3 changes

Here we list the changes that have been made to the model to comply with DCAT 3.

- dcat:Resource
  - An optional dcterms:version was added to note the version of the resource.
- dcat:Distribution
  - dcterms:hasVersion was replaced with dcat:version.

2.3 Resource metadata ontology updates

The main changes made to the resource metadata ontology were the following:

- ejprd:VPQueryable was removed as it is no longer deemed necessary.
- Classes for representing Beacon services were added, namely ejprd:VPBeacon2_Individuals and ejprd:VPBeacon2_Catalog. Both of these are subclasses of edam:operation_004 (Operation).
- An ejprd:personalData property was added for GDPR purposes.

3. Change request governance

As the EJP RD VP is becoming more mature and resources are being onboarded, it has become clear that we need to give careful consideration to change requests. To address this a governance workshop was held in the Netherlands in November 2023. In this workshop a technical committee was established to review the impact of change requests. The idea is that for technical decisions, the technical committee will review the decision and then vote on whether it will be implemented or how to proceed.

In line with this recommendation from the governance workshop, we have implemented a governance process for making changes to the resource metadata model and the resource metadata ontology.

For every change to the resource metadata model, even for very small changes, one or more GitHub issues need to be opened. This ensures full transparency of the requirements for all stakeholders.

Once a ticket is opened, an **awaiting decision** status should be assigned. This means that the issue needs to be discussed by the EJP RD technical committee to decide on whether this request should be implemented or not. If this is to be implemented, additional information can be added to the issue as to how this request should or could be implemented. Getting a request ratified is an essential step because changes to the resource metadata model has consequences for how resources can be onboarded, resources that are already onboarded, as well as for downstream systems like the FAIR data point (FDP), the query mechanism and the VP. It is these stakeholders that form part of the technical committee who are responsible for ratifying requests.

Once a request has been ratified, the status of the issue should be changed to **ratified**. Only once a request has been ratified should modellers start to implement the change. Once work on implementing a change has started, the issue status should be changed to **change in progress**. This may seem like an unnecessary status change, but its value is in communicating to stakeholders what is happening. This is particularly important in the following 2 cases:

1. There is a long delay between when a request has been ratified and when working on implementing the change is started.
2. Sometimes, despite careful consideration by the technical committee, a change that has been ratified, can be reverted. This happens when a consequence that was not immediately apparent, becomes apparent through additional information. In this case an issue with status **ratified** or **change in progress** can change back to **awaiting decision**.

In some cases a change request can be declined. This can happen when what is requested can already be achieved by using the model as-is, or if the change is not deemed to be well aligned with the objectives of the EJP RD. In this case the issue status should be changed to **declined**, ideally with a comment stating the reason for declining the request. Then the issue should be closed.

Once a change has been completed and pushed to the GitHub repository, the status of the issue should be changed to change **completed**. This step is important to ensure that the issue is indeed completed and that it was not accidentally closed. Completion here refers to changes to the model and not to changes to downstream systems. Finally, the issue can be closed.

This process is also detailed on the GitHub Wiki. Currently this process is executed using GitHub. The advantage of using the mechanisms of GitHub is that GitHub is publicly accessible and allows for asking specific people to give input. An example where this process is applied can be seen is the Biobanks metadata with MIABIS structure GitHub pull request.

4. Next steps and recommendation

There is a need for continuous alignment with Europe-wide updates to the DCAT application profiles (DCAT AP); in specific alignment with the DCAT-AP health extension (currently in development), is of importance. Close alignment with DCAT AP health extension will ensure FAIRification of data and services across a wider number of health care providers and research infrastructures.

The technical committee will be responsible for the continued alignment of the existing resource metadata model and resource metadata ontology with DCAT-AP, as well as any change requests.

Glossary

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

DCAT-AP: DCAT Application Profile is mechanism for adding essential features necessary for a specific domain or context.

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

FAIR: It is a set of guiding principles to make data and services findable, accessible, interoperable and reusable. See https://www.nature.com/articles/sdata201618

FDP: FAIR Data Point is a metadata service that provides access to metadata following the FAIR principles. See https://specs.fairdatapoint.org/

ODRL: Open Digital Rights Language is a policy expression language that provides a flexible and interoperable information model, vocabulary, and encoding mechanisms for representing statements about the usage of content and services. See https://www.w3.org/TR/odrl-model/

Beacon-2 Compliant API Specifications: Two endpoint specifications designed in the context of the EJP RD Virtual Platform to support safe resource-level and record-level querying of resources within the VP network. See https://github.com/ejp-rd-vp/vp-api-specs