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 11.4
Fourth 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 48

Dissemination level:
Public

Table of Contents

1. Introduction 3
2. Approach 4
3. Resource Metadata Model Updates 5
3.1. Resource metadata updates 5
3.2. Resource Metadata Template and Guidelines 8
4. Next Steps 9
5. Glossary 10

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, and use data for various purposes, ranging from discovering data and samples to analysing data.

This deliverable reports on progress of the fourth year (Y4) of the ongoing work pertaining to subtasks 11.1.1 and 11.1.2 of Work Package 11 which is to develop a resource metadata model and ontological model to support the EJP RD VP. In this section we first provide a summary of our work in previous years (Year 1-3), then we list the key updates for Y4 and finally we describe the structure of the rest of this document.

The work we conducted in previous years is summarised next. 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 Y4 we extended this work though the following outputs:

- We have extended the model to
    - distinguish between datasets that are discoverable and datasets that are both discoverable and queryable,
    - note conformance of a dataset to an ontology, and
    - provide links to relevant access rights information.
- One of the ways in which a resource can be onboarded to the EJP VP is via the use of a FAIR Data Point (FDP). See the EJP RD Virtual Platform: Resources Onboarding Manual for details. Both the FDP specification and the resource metadata model are based on DCAT version 2. However, we have noted some differences in these models that hindered integration and for this reason the resource metadata model has been updated to align more closely with the FDP specification.
- An FDP enables automated onboarding of resources. However, for some users a manual means to onboard their resources is desired. For this reason, we provide a spreadsheet that resource provider can fill in with the details of their resources from which an FDP can be generated.
- The spreadsheet to generate an FDP is non-trivial to fill in. For this reason, we provide detailed documentation to guide users in filling in the spreadsheet.

The rest of this report is structured as follows. In the next section we discuss the approach we followed in creating the outputs for this deliverable. Section 3 discusses in detail the updates

to resource the metadata model and related ontological model. Section 4 gives our next steps planned for Y5.

2. Approach

The work of Year 4 builds on the work that has been done in years 1-3 in which we have been able to provide a level of continuity that is impressive given the number of changes to members of this team over the last 4 years. Updates to the resource metadata model and ontological model has been informed based on workshops, meetings with resources, collaboration with EJPRD colleagues and various EJP RD project meetings.

3. Resource Metadata Model Updates

3.1. Resource metadata updates

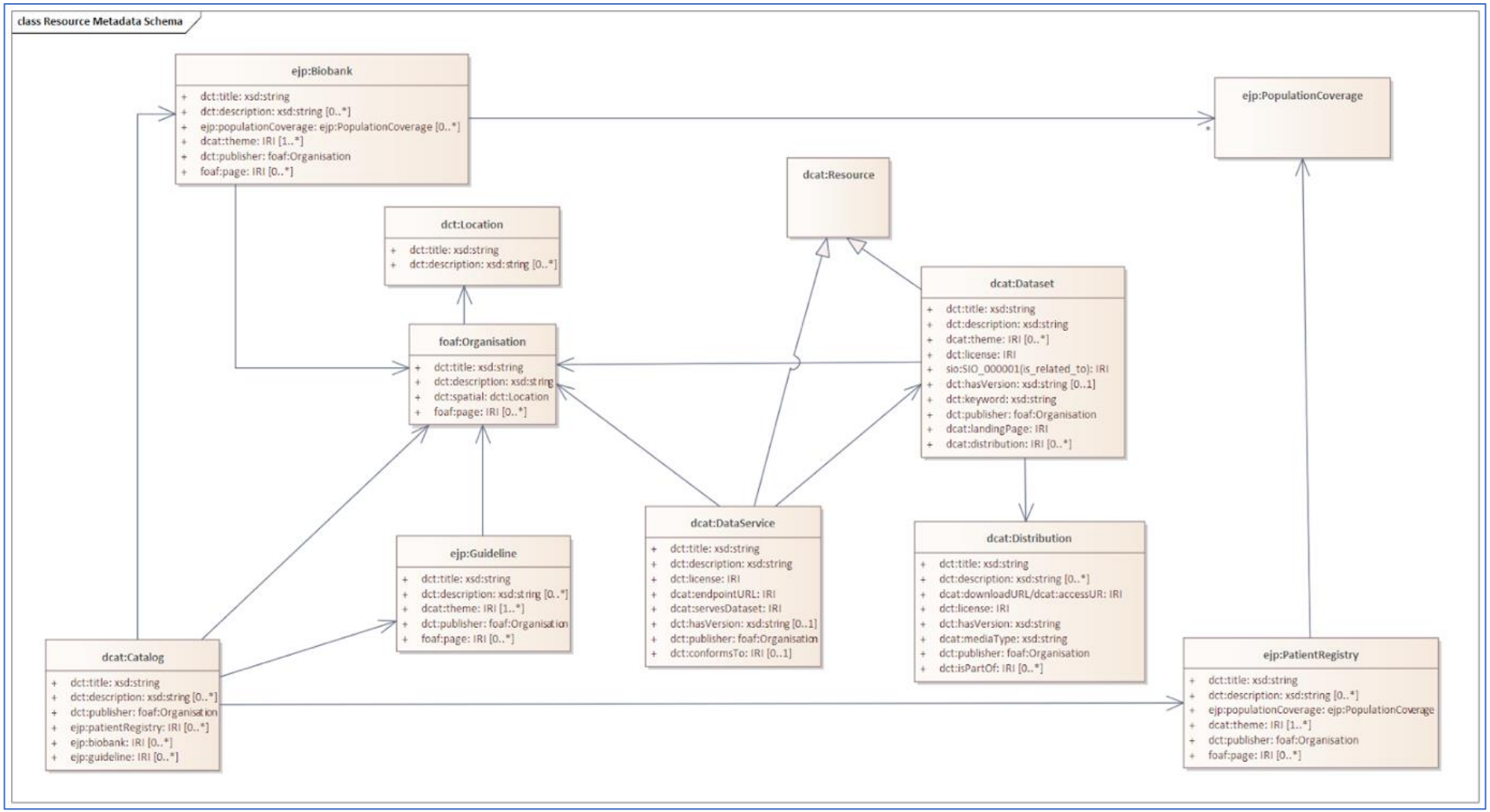

Figure1. Year 3 resource metadata model

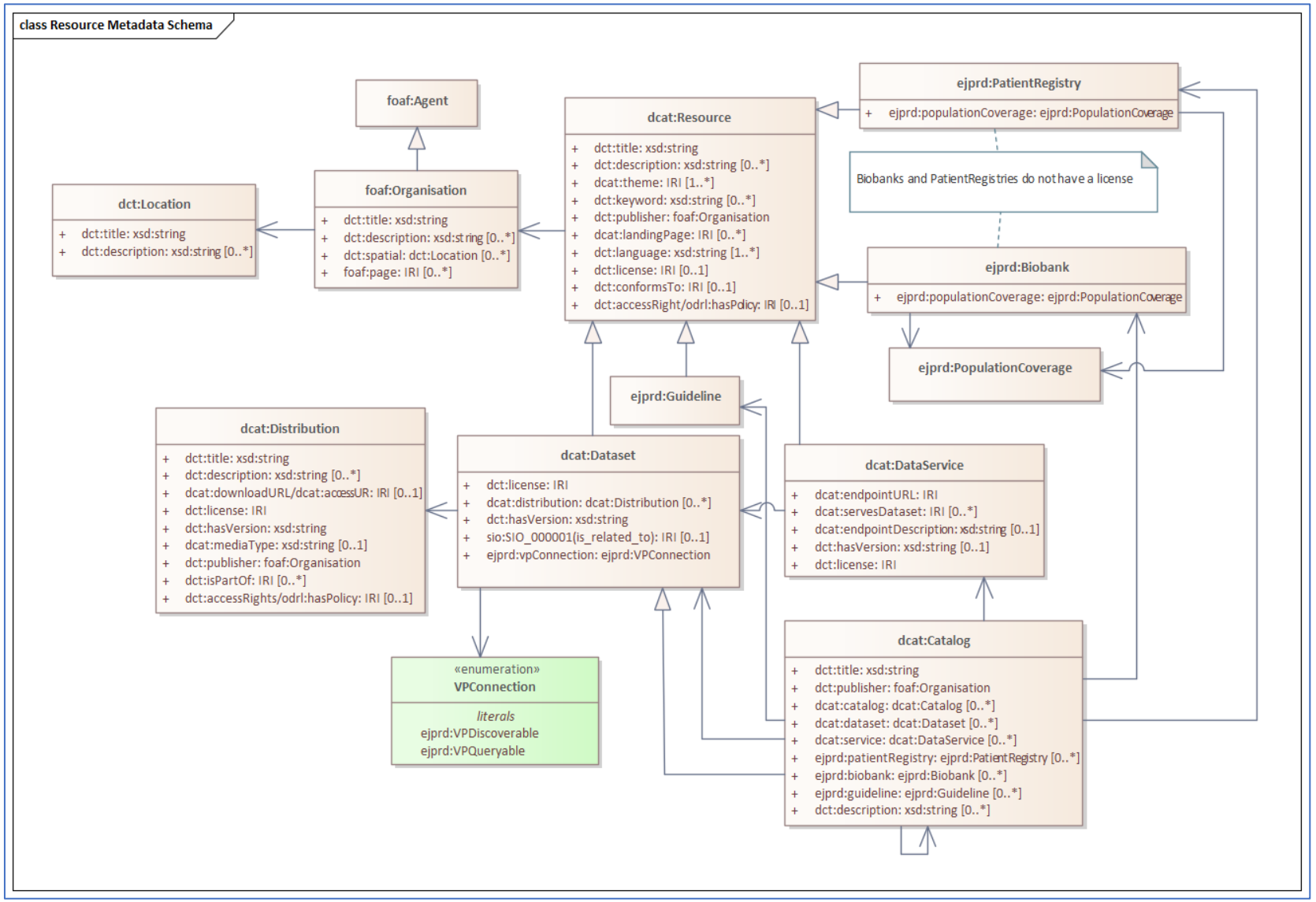

Figure 2. Year 4 resource metadata model. https://github.com/ejp-rd-vp/resource-metadata-schema

In this section we describe changes to the model to address the evolving requirements of the EJPRD VP and to align more closely with the FDP specification. The effect of these changes can be seen in comparing the Year 3 resource metadata model (Figure 1) with the Year 4 resource metadata model (Figure 2).

3.1.1 Extensions to the model and related ontological updates

In Year 4 initial steps have been taken to represent VP connection levels, data conformance and access rights. These changes are detailed here.

VP connection levels

Based on initial discussions the model has been extended for datasets to have a "ejprd:vpConnection" property of type "ejprd:VPConnection" with the possible values "VPDiscoverable" and "VPQueryable". This maps to the EJPRD VP notion of connection levels as follows:

- *Metadata discovery*. This is the minimal connection level assumed for any resource connected to the VP. As such it is not represented explicitly in the model.
- The *Data discovery* connection level is represented in the resource metadata model as VPDiscoverable.
- The *Data querying* is represented as VPQueryable in resource metadata model.

The EJPRD ontological model has been extended using the Data Use Ontology (DUO) and in specific the "data use modifier" concept has been extended with the EJPRD VP connection levels (see Figure 3) since this is a concept that indicates addition conditions for use.

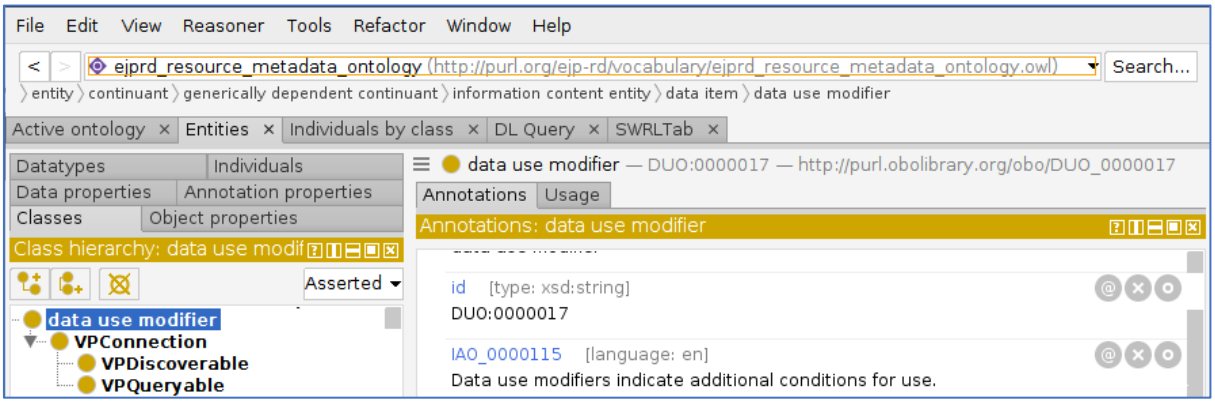

Figure 3: Ontological definition of VPConnection

Dataset conformance

The EJP RD brings together a variety of rare disease resources and these different resources use different standards or ontologies in the definition of their data. As such we have introduced the "dct:conformsTo" property from Dublin Core which is used to establish a standard a resource conforms to.

Access rights

To cater for the different access rights that different EJPRD resources can have, we added an "dct:accessRights" property from Dublin Core to provide details about access or

restrictions based on privacy, security, or other policies, and an "odrl:hasPolicy" property if this information is defined in a way that is compliant to ODRL.

3.1.2 Alignment with FDP specification

In our collaboration with the FDP developers, it became clear that the FDP specification and resource metadata models have differences despite both being based on DCAT version 2. Here we list and describe our amendments to align the resource metadata specification more closely that of the FDP specification:

- DataService:
  - We have added a "dcat:endpointDescription" property to add zero or more strings of text for describing a data service.
  - We changed the cardinality of the "dcat:servesDataset" property from required 1 to zero or more.
- Dataset:
  - In our model we use the "dct:keyword" from Dublin Core instead of the "dcat:keyword" from DCAT version 2, which we corrected.
  - We did not provide a way to define the language that is used for a dataset. Therefore we added a "dct:language" property in accordance with DCAT version 2 and the FDP specification.
- Distribution:
  - Specifying one of the "dcat:downloadURL" or "dcat:accessURL" properties was required in our model, for which we changed the cardinality be optional or one.
  - The FDP specification only allows for at most a single mediatype to be associated with a distribution. Our model allowed for multiple media types which we revised to adhere to the FDP specification.
- Resource, Guideline, PatientRegistry, Biobank: In our model the landing page for a resource was defined using "foaf:page" instead of "dcat:landingPage", which we corrected. This resulted in similar changes for Guideline, PatientRegistry and Biobank.

3.2. Resource Metadata Template and Guidelines

An update from Year 4 is that we provided a spreadsheet to allow for manual creation of FDP based on the import of the spreadsheet data to an FDP. Since this spreadsheet caters for the richness of data represented by the resource metadata model, the spreadsheet can be difficult to populate. For this reason, we provide a detailed activity diagram (see Figure 4) and a description of the steps for populating the spreadsheet.

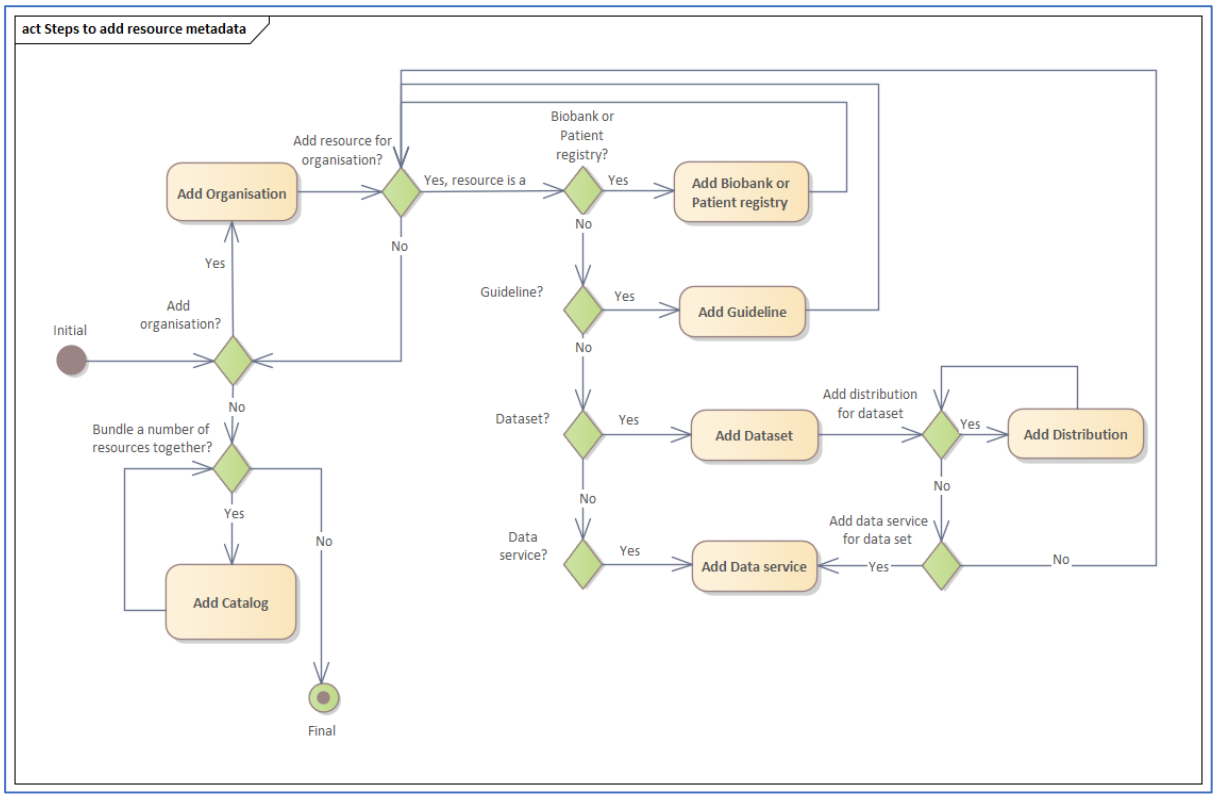

Figure 4. Activity diagram as guidance for population of spreadsheet

4. Next Steps

In Year 5 we will continue our work of previous years in aligning the resource metadata model and the resource metadata ontology in support of the evolving needs of the EJPRD VP. This will include the following tasks.

- The VP connection level is currently modelled at Dataset level. However, this is a concern that is relevant to all resources and hence is perhaps better suited to be modelled as such. Moreover, the VP connection level should also cater for federated analysis in accordance with the EJP RD Virtual Platform: Resources Onboarding Manual.
- We will be extending our support for catalogues of biobanks, patient registries and guidelines to include support for animal models and cell lines.
- We will extend the resource metadata model to support or extend support for the developing needs around
    - quality and sustainability,
    - GDPR
    - Common Conditions of use Elements (CCE) and Digital Use Condition (DUC).

5. Glossary

CCE: Common Conditions of use Elements. See https://duc.le.ac.uk/Learn/index

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

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

DUC: Digital Use Conditions. See https://duc.le.ac.uk/Learn/index

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/