Using DCAT-AP for research data

Authors Andrea Perego, Anders Friis-Christensen, Lorenzino Vaccari, Chrisa Tsinaraki

Affiliation European Commission, Joint Research Centre (JRC) (https://ec.europa.eu/jrc/)

Abstract

This paper outlines a set of cross-domain requirements for the documentation of scientific data, identified during the development of the corporate data catalogue of the European Commission's Joint Research Centre (JRC).

In particular, we illustrate how we have extended the *DCAT application profile for European data portals* (DCAT-AP) to accomodate requirements for scientific datasets, and we discuss a number of issues still to be addressed.

Workshop	*Smart Descriptions & Smarter Vocabularies* (SDSVoc). Amsterdam, 30 Nov - 1 Dec 2016 (https://www.w3.org/2016/11/sdsvoc/).
URL	https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_27
Last modified	5 Dec 2016

Disclaimer: The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

Introduction

The overall mission of the Joint Research Centre of the European Commission (JRC) (https://ec.europa.eu/jrc/) is to support EU policies with independent evidence throughout the whole policy life-cycle. The activities of the JRC span many different research areas, that address and ensure a healthy and safe environment, secure energy supplies, sustainable mobility and consumer health and safety. Thus, the JRC is a multidisciplinary research organisation and the diversity of its research activities poses challenges in the management of data, since different scientific disciplines have their own traditions, standards and best-practices on how to manage and disseminate research information.

In order to provide a basis for better management of data, in 2014 the JRC developed a corporate data policy [13], driven by the need of transparency in the policy development cycle, and to facilitate open access to research data, in line with the general Open Data trend.

As part of the activities concerning the implementation of the JRC Data Policy, JRC has set up a corporate catalogue (http://data.jrc.ec.europa.eu/), which is meant to be a single point of access to data produced and/or maintained by JRC. In this context, metadata play a fundamental role, and are meant to address a number of requirements, which include (a) ensuring data documentation and inventory, (b) enabling data discovery and (c) publishing metadata on other catalogues operated by EU institutions and bodies - in particular, the EU Open Data Portal (http://data.europa.eu/euodp/).

The multidisciplinary nature of JRC datasets makes it difficult to define a common metadata schema able to fit all requirements. Therefore, the design of the JRC metadata schema is following a modular approach consisting of a *core* profile, defining the elements that should be common to all metadata records, and a set of domain-specific *extensions*.

The reference metadata standard used is the *DCAT application profile for European data portals* (DCAT-AP) [3] (the *de facto* EU standard metadata interchange format), and the related domain-specific extensions - namely, GeoDCAT-AP [8] (for geospatial metadata) and StatDCAT-AP (for statistical metadata) [16].

The core profile of JRC metadata is however not using DCAT-AP *as is*, but it complements it with a number of metadata elements that have been identified as most relevant across scientific domains, and which are required in order to support data citation.

The following sections provide a summary of the adopted solutions, and discuss a number of issues still to be addressed.

Metadata elements relevant for scientific data

The most common, cross-domain requirements we identified for JRC data are following ones:

1. Ability to indicate dataset authors.
2. Ability to describe data lineage.
3. Ability to give potential data consumers information on how to use the data ("usage notes").
4. Ability to link to scientific publications about a dataset.
5. Ability to link to input data (i.e., data used to create a dataset).

Points (2) (data lineage) and (5) (input data) are already supported by DCAT-AP via, respectively, `dct:provenance` (http://purl.org/dc/terms/#terms-provenance) and `dct:source` (http://purl.org/dc/terms/#terms-source). For the other ones, our approach is as follows:

- Dataset authors: `dct:creator` (http://purl.org/dc/terms/#terms-creator).
- Usage notes: `vann:usageNote` (http://vocab.org/vann/#usageNote).
- Related (scientific) publications: `dct:isReferencedBy` (http://purl.org/dc/terms/#terms-isReferencedBy).

Information on dataset authors, input data and publications provide a first, simplified, picture of the "context" of a dataset, that can be used for multiple purposes. For instance, to filter datasets based on their authors or input data, to identify the most used datasets, the use cases where a dataset has been used. However, to support this effectively, a key requirement is the ability to rely on persistent identifiers.

In this perspective, we have integrated ORCIDs (http://orcid.org/), whenever available, in dataset authors metadata, whereas all the dataset published on the JRC Data Catalogue are assigned a persistent URI from the URI registry of EU institutions and bodies (http://data.europa.eu/) operated by the Publications Office of the EU. This will ensure the long-term availability of metadata records also in case the existing infrastructure is migrated.

Such an approach ensures the ability of linking resources maintained at the corporate level, but of course this does not apply to external ones (as it happens frequently for input data) not associated with a persistent identifier.

Data citation

DataCite (https://www.datacite.org/) is an international initiative meant to enable citation for scientific datasets. To achieve this, DataCite operates a metadata infrastructure, following the same approach used by CrossRef (http://www.crossref.org/) for scientific publications. As such, the DataCite infrastructure is responsible for issuing persistent identifiers (in particular, DOIs) for datasets, and for registering dataset metadata. Such metadata are to be provided according to the DataCite metadata schema – which is basically an extension to the one used for DOI records.

Since DataCite is currently the *de facto* standard for data citation, we started by carrying out a preliminary study concerning the mapping of DataCite with DCAT-AP, and by developing an experimental, XSLT-based, implementation of the defined transformation rules [2].

Based on this work, we recognise that what needs to be added in DCAT-AP for data citation purposes is basically only one "field" (already mentioned above), namely, "data authors".

Other issues concern mainly (a) identifiers (DOIs, ORCIDs, ISNIs, ISSNs, etc.) and (b) agent roles, which are elaborated in the following sections.

Identifiers

The requirements are basically the following ones:

- DataCite requires the dataset identifier to be a DOI.
- DataCite distinguishes between primary and secondary identifiers.
- DataCite models the "type" of identifier (DOIs, ORCIDs, ISNIs, ISSNs, etc.).

DCAT-AP already provides a mechanism to model primary and secondary identifiers, as well as the identifier type. More precisely:

- Property `dct:identifier` (http://purl.org/dc/terms/#terms-identifier) is used to model primary identifiers.
- Property `adms:identifier` (https://www.w3.org/TR/vocab-adms/#adms-identifier) is used to model secondary/alternative identifiers.
- Class `adms:Identifier` (https://www.w3.org/TR/vocab-adms/#identifier) allows the specification of information about the identifier - identifier scheme included. More precisely, the identifier is specified by using property `skos:notation` (https://www.w3.org/TR/skos-reference/#notations), typed with the URI of one of the members of the DataCite Resource Identifier Scheme (http://www.sparontologies.net/ontologies/datacite/source.html#d4e641).

Such solutions are basically reflecting the DataCite approach to model identifiers. However, it is questionable whether they fit requirements for different or more general scenarios. In particular, the issue is that identifiers modelled in this way are of no use for effectively linking the relevant resources. For this purpose, it would be desirable to promote the encoding of identifiers as HTTP URIs, whenever possible. This is the case, e.g., for ORCIDs, ISNIs, and DOIs. Notably, some of the relevant identifier services already offer the ability to retrieve machine-readable metadata by dereferencing URIs (e.g., this applies to ORCIDs and DOIs). Finally, about the ability to modelling differently primary and secondary/alternative identifiers: the resource URI can denote the primary identifier, whereas URIs corresponding to alternative identifiers can be specified by using `owl:sameAs` (https://www.w3.org/TR/owl-ref/#sameAs-def).

Another issue concerns the actual benefits of modelling identifiers with a specific class. E.g., `adms:Identifier` was meant to enable the specification of the identifier scheme agency and the identifier issue date, following the conceptual definition of the UN/CEFACT Identifier class [1]. However, if only the identifier scheme is required, the question is whether it would be possible to simplify the current representation. Possible options include the following:

1. Modelling identifier schemes as datatypes. In such a case, it would be possible to use just `dct:identifier`, typed with the relevant identifier scheme data type.
2. Defining specific properties for each identifier scheme (as sub-properties of `dct:identifier`). Notably, some of these properties are already defined in existing vocabularies - e.g., `bibo:doi` (http://bibliographic-ontology.org/) -, and therefore they can be re-used.

Based on what said above, the possible solutions, alternative to the current one, can be summarised as follows:

1. Encode identifiers as (HTTP) URIs, whenever possible (DOIs, ORCIDs, etc.), using `owl:sameAs` for (HTTP) URIs concerning secondary/alternative identifiers.
2. Model identifiers with `dct:identifier`, typed with the relevant identifier scheme data type, or with specific subproperties of `dct:identifier`.
3. If the ability of denoting identifiers as secondary/alternative is a requirement, use `adms:identifier`.

It is worth noting that these three options are not mutually exclusive.

Agent roles

DataCite supports three main types of agent roles, namely, author, publisher, and contributor. The last can be further specialised by specifying a contributor "type". At the time of writing this paper, DataCite supports 22 contributor types, including, e.g., "contact person", "data curator", "distributor", "editor", "producer", "rights holder", "other".

This situation is not different from other metadata standards. E.g., ISO 19115 [12], the standard for geospatial metadata, originally included 11 agent roles - a number that, in the latest version of this standard, has increased to 20.

The issue arises when trying to share and re-use these metadata records, since this information might be lost, unless it is mapped consistently. On the other hand, the question is also if such information could be actually relevant - e.g., it might important to preserve information only the of "key" roles, as dataset creator, publisher and contact point.

The current version of DCAT-AP (v1.1) supports only two agent roles, namely, data publisher and contact point. GeoDCAT-AP includes other two ones - namely, dataset creator and rights holder - but, in addition, it defines a mechanism to model all the ISO 19115 roles, making use of the W3C PROV Ontology [15]. It is worth noting that this feature is only supported in the full GeoDCAT-AP profile, which is meant to provide a complete representation of the metadata elements defined in the core profile of ISO 19115 and in the INSPIRE metadata specification [11].

An example is provided by the following code snippet:

EXAMPLE

```
a:Dataset a dcat:Dataset;
  prov:qualifiedAttribution [ a prov:Attribution ;
# The agent role, as per ISO 19115
    dct:type <http://inspire.ec.europa.eu/metadata-codelist/ResponsiblePartyRol
e/owner> ;
# The agent playing that role
    prov:agent [ a foaf:Organization ;
      foaf:name "European Union"@en ] ] .
```

Example of a GeoDCAT-AP PROV-based representation of an agent role.

The PROV-based solution defined in GeoDCAT-AP has the advantage of being domain-independent - e.g., it could be re-used also to model DataCite agent roles. Moreover, it provides the ability to attach additional information (e.g., during which timeframe a given agent played a given role). However, it has two drawbacks:

1. It is overly complex, compared with the use of *simple* role properties - as `dct:creator`, `dct:publisher`, `dct:contributor`.
2. To denote the role, it makes use of URIs operated by the INSPIRE Registry (http://inspire.ec.europa.eu/registry/) for the agent roles defined in ISO 19115. As a consequence, if roles defined in another standard are used, the code list will be different, and interoperability will not be granted.

A possible solution has been discussed during the revision of DCAT-AP, based on the idea of maintaining a "role property vocabulary", that could also be used to bridge the agent roles defined in the different metadata standards. So far, this seems to be the most viable solution, at least in the framework of DCAT-AP, and related profiles. Such "role property vocabulary" could also be used as a means to prevent the inconsistent use of agent roles. This issue is more and more apparent in metadata standards supporting multiple roles, with overlapping semantics (e.g., the difference between a data distributor and a data publisher is not always clear). We do not have evidence of such a situation in DataCite. However, as far as geospatial metadata are concerned, a study [17] has been carried out in 2014 on the records available from the INSPIRE Geoportal (http://inspire-geoportal.ec.europa.eu/), harvested across EU Member States. The results show that, among the 11 agent roles defined in ISO 19115, the one most used was "point of contact", followed by "owner", whereas the collected statistics, aggregated by country, suggest that roles “custodian”, “distributor”, “publisher” and “resource provider” were used to denote the same role.

A consistent use of agent roles is crucial to enable metadata interoperability. If not ensured, preserving this information would be useless or even counterproductive. In such a case, identifying a minimal set of unambiguous roles and promote their use would be preferable - e.g., those already supported by DCAT-AP, plus a few ones.

Modelling service/API-based data access

This issue concerns dataset distribution that are made accessible via services and APIs. Examples include SPARQL endpoints, as well as the download and view services used for geospatial data. In such cases, users, who expect to get to the actual "data", are instead returned an API query interface, usually meant to be used by software agents. On the other hand, software agents are not provided enough information on how to access the data via the target service / API. Finally, an additional issue is that a service / API may provide access to more than one dataset. As a consequence, users (as well as software agents) do not know how to get access to the subset of relevant data accessible via a service. Requirements to address this issue are basically two:

1. Denote distributions as pointing to a service / API, and not directly to the actual data.
2. Provide a description of the API / service interface, along with the relevant query parameters, that can be directly used by software agents - either to access the data, or to make transparent data access to end users.

We are currently addressing point (1) by associating with distributions the following information:

- Whether the access / download URL of a distribution points to data or to a service / API (`dct:type`).
- In the latter case, we include the specification the service/API conforms to (`dct:conformsTo`).

An example is provided by the following code snippet. Here, the distribution's access URL points to service, implemented by using the WMS standard of the *Open Geospatial Consortium* (OGC) (http://www.opengeospatial.org/):

EXAMPLE

```
a:Dataset a dcat:Dataset;
  dcat:distribution [ a dcat:Distribution ;
    dct:title "GMIS - WMS (9km)"@en ;
    dct:description "Web Map Service (WMS) - GetCapabilities"@en ;
    dct:license <http://publications.europa.eu/resource/authority/licence/COM_R
EUSE> ;
    dcat:accessURL <http://gmis.jrc.ec.europa.eu/webservices/9km/wms/meris/?dat
aset=kd490> ;
# The distribution points to a service
    dct:type <http://publications.europa.eu/resource/authority/distribution-typ
e/WEB_SERVICE> ;
# The service conforms to the WMS specification
    dct:conformsTo <http://www.opengis.net/def/serviceType/ogc/wms> ] .
```

Example of a distribution pointing to a WMS service.

As far as point (2) is concerned (i.e., provide a description of the API / service interface), we are considering a number of possible options. Currently, the one that seems promising is the proposal, developed in the framework of the *DCAT-AP implementation guidelines* [4], to describe a service/API by using an OpenSearch document [14] - see issue *DT2: Service-based data access* (https://joinup.ec.europa.eu/asset/dcat_application_profile/issue/dt2-service-based-data-access) on JoinUp.

Additional requirements

We include below some additional requirements we are investigating, and that we cannot elaborate further due to space constraints.

Detailed specification of data provenance
This is meant to provide a detailed representation of the "data context", including all the entities and activities involved in the data life-cycle. The objectives include (a) data reproducibility and (b) ability

to track the usage of data as well as the models used for their creation.

Modelling and using data quality assessments

This is not limited to data, but concerns metadata as well, and it potentially includes the integration of users' feedback as part of the meta/data management life-cycle. On this topic, we are considering approaches making use of vocabularies as the W3C PROV Ontology [15], the Data Quality Vocabulary (DQV) [6] and the Dataset Usage Vocabulary (DUV) [7].

Optimising metadata publication on the Web

One of the use cases concerns increasing visibility and discoverability of JRC data on the Web, also via search engines. It basically relies on the use of mechanisms as HTML+RDFa [10], but it includes as well mapping exercises with popular Web vocabularies, as Schema.org - see, e.g., [5].

Some of these issues are discussed in a separated paper [9], to which we refer the reader.

References

[1] *Core Components Data Type Catalogue. Version 3.1* (2011) United Nations Centre for Trade Facilitation and Electronic Business (UN/CEFACT) http://www.unece.org/fileadmin /DAM/cefact/codesfortrade/CCTS/CCTS-DTCatalogueVersion3p1.pdf

[2] *DataCite to DCAT-AP Mapping* (2016) European Commission, Joint Research Centre (JRC) https://webgate.ec.europa.eu/CITnet/stash/projects/ODCKAN/repos/datacite-to-dcat-ap/

[3] *DCAT application profile for data portals in Europe* (2015) EU ISA Programme (ISA²) https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe

[4] *DCAT application profile implementation guidelines* (2016) EU ISA Programme (ISA²) https://joinup.ec.europa.eu/solution/dcat-application-profile-implementation-guidelines

[5] *DCAT-AP to Schema.org Mapping* (2016) European Commission, Joint Research Centre (JRC) https://webgate.ec.europa.eu/CITnet/stash/projects/ODCKAN/repos/dcat-ap-to-schema.org/

[6] *Data on the Web Best Practices: Data Quality Vocabulary* (2016) World Wide Web Consortium (W3C) https://www.w3.org/TR/vocab-dqv/

[7] *Data on the Web Best Practices: Dataset Usage Vocabulary* (2016) World Wide Web Consortium (W3C) https://www.w3.org/TR/vocab-duv/

[8] *GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe* (2016) EU ISA Programme (ISA²) https://joinup.ec.europa.eu/node/139283

[9] *GeoDCAT-AP: Use cases and open issues* (2016) European Commission, Joint Research Centre (JRC) https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_25

[10] *HTML+RDFa 1.1 - Second Edition: Support for RDFa in HTML4 and HTML5* (2015) World Wide Web Consortium (W3C) https://www.w3.org/TR/html-rdfa/

[11] *INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119. Version 3.1* (2013) European Commission, Joint Research Centre (JRC) http://inspire.ec.europa.eu/documents/Metadata/MD_IR_and_ISO_20131029.pdf

[12] *ISO 19115:2003: Geographic information -- Metadata* (2003) International Organization for Standardization (ISO) https://www.iso.org/standard/26020.html

[13] *JRC Data Policy* (2015) European Commission, Joint Research Centre (JRC) doi:10.2788/607378

[14] *OpenSearch* (2016) OpenSearch.org http://www.opensearch.org/

[15] *PROV-O: The PROV Ontology* (2013) World Wide Web Consortium (W3C) https://www.w3.org/TR/prov-o/

[16] *StatDCAT application profile for data portals in Europe* (2016) EU ISA Programme (ISA²) https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/

[17] *Use of responsible party roles in INSPIRE metadata* (2014) European Commission, Joint Research Centre (JRC) https://ies-svn.jrc.ec.europa.eu/projects/metadata /wiki/Use_of_responsible_party_roles