DCAT-AP implementation guidelines

Working Group meeting

17 March 2017

OPENING, AGENDA, TOUR DE TABLE

Agenda

1. Opening, agenda, tour de table
2. Objectives
3. Technical guidelines
4. Non-technical guidelines
5. Future work
6. Ontologies
7. Closing

TOUR DE TABLE

OBJECTIVES

Objectives

- Present the outcomes of our interviews with several implementers of DCAT-AP.
- Discuss the issues identified and any further comments from the interviewees.
- Define next steps towards finalising the DCAT-AP implementation guidelines.

Discussion of proposed guidelines

The guidelines are divided in two groups:

- Technical guidelines, e.g.:
    - How to use accessURL and downloadURL?
    - How to refer to media types within ZIP files?

- Non-technical guidelines, e.g.:
    - Which processes and tools are being used or can be used to guarantee the quality of metadata?

TECHNICAL GUIDELINES

European Commission

1. How to use accessURL and downloadURL?

Link: https://joinup.ec.europa.eu/node/156370

Definitions:

- *accessURL*: this property contains a URL that gives access to a Distribution of the Dataset. The resource at the access URL may contain information about how to get the Dataset.

- *downloadURL*: this property contains a URL that is a direct link to a downloadable file in a given format.

Issue: The relationship between downloadURL and accessURL may create duplication of information in certain cases due to the fact that accessURL is mandatory.

ISA²

European Commission

1. How to use accessURL and downloadURL?

Guideline:

- dcat:accessURL: direct access to a file or to a page containing further instructions;
- dcat:accessURL is mandatory and guarantees the existence of descriptions for the distributions.
- dcat:downloadURL: direct link to a file. It allows software programs to use the link to get access to the file;
- If direct download access can be provided, the URL of the data should be duplicated in both accessURL and downloadURL.
- Currently there is no standardised way to refer to service-based data access. dcat:accessURL can be used to link to the landing page of the service that gives access to the data.

ISA²

1. How to use accessURL and downloadURL?

Example: based on the Nobel Prize catalogue available via http://www.nobelprize.org/datasets/dcat

```
<rdf:Description rdf:about="https://dcat-editor.com/store/17/resource/6">
    <ns1:downloadURL rdf:resource="http://api.nobelprize.org/v1/laureate.csv"/>
    <rdf:type rdf:resource="http://www.w3.org/ns/dcat#Distribution"/>
    <ns1:accessURL rdf:resource="http://api.nobelprize.org/v1/laureate.csv"/>
</rdf:Description>
```

ISA²

2. How are publisher and contact point modelled?

Link: https://joinup.ec.europa.eu/node/149022

Definitions:

- *Publisher:* the real-world entity of class foaf:Agent. An entity (organisation) responsible for making the Dataset available.

- *Contact point:* contact information that can be used for sending comments about the Dataset. The contact information of class vCard:Kind associated to an organisation, not the organisation itself.

Issue: the distinction from DCAT and DCAT-AP between Publisher and Contact point is often misinterpreted.

Guideline:

- Semantic difference at the conceptual level: foaf:Organization is a real-world entity, vcard:Organization is like the information on a business card.

- It is important to differentiate the two: Publisher is necessary to identify the entity and Contact point allows any person/organisation to communicate and provide feedback.

ISA²

European Commission

2. How are publisher and contact point modelled?

Example: based on the Nobel Prize catalogue available via
http://www.nobelprize.org/datasets/dcat

```
<rdf:Description rdf:about="https://dcat-editor.com/store/17/resource/1">
    <foaf:mbox rdf:resource="mailto:technical@nobelprize.org"/>
    <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Agent"/>
    <foaf:homepage>http://www.nobelprize.org/</foaf:homepage>
    <foaf:name>Nobel Media</foaf:name>
    <dcterms:type rdf:resource="http://purl.org/adms/publishertype/Company"/>
</rdf:Description>
<rdf:Description rdf:about="https://dcat-editor.com/store/17/resource/2">
    <ns3:fn>Hans Mehlin</ns3:fn>
    <rdf:type rdf:resource="http://www.w3.org/2006/vcard/ns#Individual"/>
    <ns3:hasEmail rdf:resource="mailto:hans.mehlin@nobelmedia.se"/>
    <ns3:hasTelephone rdf:nodeID="_node19gmorse5x1"/>
    <ns3:hasAddress rdf:nodeID="_node19gmorse5x2"/>
</rdf:Description>
```

ISA²

3. How should dct:spatial and dct:Location be used?

Link: https://joinup.ec.europa.eu/node/156379

Definitions

- *dct:spatial*: continents, countries and places that are in MDR NAL lists, Geonames URIs must be used if a place is not in a NAL.
- *dct:Location*: a spatial region or named place. It can be represented using a controlled vocabulary or with geographic coordinates.

Issue: DCAT-AP version 1.1 lists a number of controlled vocabularies for named places but it is not the only way to express spatial coverage.

Guideline:

- As expressed in the DCAT-AP specification, dct:location can be represented using a controlled vocabulary or with geographic coordinates;
- In the case geographic coordinates are used, the Core Location Vocabulary is recommended, following the approach described in the GeoDCAT-AP specification.

ISA²

3. How should dct:spatial and dct:Location be used?

Example: based on an example of the GeoDCAT-AP specification on how to specify the spatial coverage.

```
[] dct:spatial [ a dct:Location ;
        locn:geometry "POLYGON((-10.58 70.09,34.59 70.09,34.59 34.56,-10.58 34.56, -
10.58 70.09))"^^gsp:wktLiteral ;
        locn:geometry "<gml:Envelope
srsName=\http://www/opengis.net/def/crs/DGC/1.3/CRS84\>
                <gml:lowerCorner>34.56 -10.58</gml:lowerCorner>
                <gml:upperCorner>70.09 34.59</gml:upperCorner>
                </gml:Envelope> "^^gsp:gmlLiteral ] ;
        locn:geometry
"{\"type\":\"Polygon\",\"crs\":{\"type\":\"name\",\"properties\":{\"name\":\"ur
n:ogc:def:crs:OGC:1.3:CRS84\"}}, \"coordinates\":[[[-
10.58,70.09],[34.59,70.09],[34.59,34.56],[-10.58,34.56],[-10.58,70.09]]] }"^^
https://www.iana.org/assignments/media-types/application/vnd.geo+json ].
```

ISA²

4. How could you make vocabularies de-referenceable?

Link: https://joinup.ec.europa.eu/node/149028

Issue: Publishers of data need to be able to handle the various situations in which URIs do not deliver metadata about the resource.

Guideline:

- URIs are required as values for properties that link to resources.
- There are two types of responses coming back from a HTTP GET on the URI:
  - One containing information about that individual URI;
  - Another containing a file in which information about the URI is combined with information about other URIs, for example in a schema file;
- In any case, an organisation that mints URIs should strive to keep the delay between minting URIs and resolving them as short as possible.

ISA²

4. How could you make vocabularies de-referenceable?

Example:

1st case: data.europa.eu/3rx/EUProgramme/ contains information about that particular URI.
2nd case: https://www.w3.org/ns/org#Organization contains information about the URI combined with information about other URIs

Related issue: the resolution of URIs should be maintained over time.

Guideline: European Commission has established a persistent URI policy and resolution service at http://data.europa.eu/ that aims to persistently resolve all URIs that are minted for resources maintained by the European Commission. The EU URI policy has the following characteristics: governance, URI pattern, URI redirection, service level guarantees.

ISA²

5. How should URI for organisations be implemented?

Link: https://joinup.ec.europa.eu/node/149026

Issue: URI needs to be available for the publisher to reference the organisation responsible for making the dataset available in the dataset description.

Guideline:

- Use of blank nodes is discouraged;
- In some environments, a central URI set is available or in the process of being set up, e.g. for public administrations in a country

European Commission

5. How should URI for organisations be implemented?

Example:

- An example of an official collection of URIs is the MDR *Corporate body* Name Authority List at the Publications Office of the EU. For instance, an organisation such as the European Coal and Steel community (ECSC) has the following URI: http://publications.europa.eu/resource/authority/corporate-body/ECSC

- In Belgium, there is an unofficial list KBO data based on data from the Crossroads Bank for Enterprises (BCE/KBO).

Future activities: cooperation between European and national levels could be considered to share approaches and to find common ways to assign URIs to organisations.

ISA²

6. How could an entity-ID service be provided?

Link: https://joinup.ec.europa.eu/node/149111

Issue: RDF requires every entity in the graph to have a URI in order to allow making assertions about the entity. As these URIs are created when the graph is created, the URIs are usually minted locally. In many cases these URIs are created randomly, or created by the software being used.

Guideline: A common URI pattern could be applied, following a clear structure that can be applied across applications in a country.

Example:

In Italy, a fixed pattern is used for these URIs: http://dati.gov.it/resource/(type)/(id), with (type) being "Catalogo", "Dataset", "Distribuzione", "PuntoContatto" etc., and (id) being a local string.

ISA²

7. How to express agent roles?

Link: https://joinup.ec.europa.eu/node/156382

Issue: DCAT-AP includes a single property to refer to an Agent (person or organisation) related with a dataset: **publisher**. While in some national profiles, additional Agent roles are defined (e.g. **creator**, **rights holder**).

Guideline: we found no widespread use of other common types of Agent roles. No extension of the number of Agent-related properties in DCAT-AP is currently foreseen.

Future activity: depending on the emerging needs, an extension could be considered in the future.

ISA²

8. How to refer to media types within ZIP files?

Link: https://joinup.ec.europa.eu/node/152419/

Definition:

- *dcat:mediaType* refers to the media type of the Distribution as defined in the official register of media types (MIME) managed by IANA.

  Issue: Implementers do not have the possibility to provide information about the format of files (dcat:mediaType) within a package (e.g. ZIP file)

  Guideline: The recommendation proposed by IETF should be followed, as to know to add '+zip' as suffix in the structured syntax to the IANA registry for media types d for files within a ZIP package.
  https://tools.ietf.org/html/rfc6839#section-3.6

  Example:

In the IANA registry, a text file in CSV format will have the following syntax structure: text/csv. In the case of a ZIP package, the syntax would become: text/csv+zip.

ISA²

NON-TECHNICAL GUIDELINES

Products and tools

- Many implementations based on CKAN + DCAT-AP extension
- Other products used: DKAN, Socrata, The Data Tank, local developments
- Further tools, e.g. input of metadata, mostly developed locally; validation often based on DCAT-AP validator.
- Most tools are Open Source
- ISA^2 willing to talk to vendors to help implementation
- WG members are invited to share tools on Joinup at https://joinup.ec.europa.eu/node/150350/

Legal and organisational challenges

- The selection and expression of licences is a challenge
  - EDP's licensing assistant is helpful https://www.europeandataportal.eu/en/content/show-license but in some cases national policies require other specific licences
  - It is usually not specified how attribution needs to be done
- DCAT-AP national profiles are usually not enforced through legislation, but voluntary adoption is encouraged
- Documentation is now mainly technical but more management-level information about objectives and benefits would be useful

ISA²

9. Which processes and tools could be used to manage the quality of metadata?

Link: https://joinup.ec.europa.eu/node/159706

Issue: Difficulties to manage appropriately the quality of the metadata.

Guideline (1/2):

Data management processes and data governance to make sure that the right person provides the data at the right moment in the process.

Two categories of metadata quality:

- Metadata should **conform to DCAT-AP** (mandatory elements, prescribed data type, controlled vocabularies etc.);
  - o Check by user interface for metadata creation with feedback in real-time to the creator.
  - o For harvesting metadata, a validator could be used (e.g. DCAT-AP validator)

ISA²

European Commission

9. Which processes and tools could be used to manage the quality of metadata?

Guideline (2/2):

- Metadata values should **accurately reflect the content** of the dataset.
  - Detailed guidelines such as EDP's Gold Book for Data Publishers or local guides to help dataset publishers to understand the elements of DCAT-AP.
  - Overall statistics on metadata quality such as the Metadata Quality Dashboard proposed by the EDP on a weekly basis.

Example:

Metadata for datasets should be created as part of the workflow as close to the source as possible, helping metadata creators with good guidelines and user interfaces that include drop-down lists and immediate feedback.

Future activities:

Consideration of emerging approaches for specification and automated execution of quality checking, e.g. based on SPARQL or the current work at W3C on the Shapes Constraint Language (SHACL, https://www.w3.org/TR/shacl/)

ISA²

Guidelines repository

- European-level guidelines are available from the European Data Portal: Goldbook for Data Providers, https://www.europeandataportal.eu/en/providing-data/goldbook
- It could be useful to share national guidelines that others can look at for inspiration
- All languages are accepted!

Open invitation to share your guidelines with the community on the Joinup page dedicated to this:

https://joinup.ec.europa.eu/node/159705

FUTURE WORK

Issues for future work on DCAT-AP

- Common machine-readable licences with clear specification of how to attribute
- Integration of GeoDCAT-AP and StatDCAT-AP with DCAT-AP
- Standard expression of service-based data access (incl. APIs), on the basis of proposed work at W3C
- Statistical analysis on the practical use of the defined properties in DCAT-AP
- Translation of some of the content on Joinup

Next steps

1. Feedback from Working Group, until mid April
2. Finalisation of the guidelines: end of April
3. Publication on Joinup: early May
4. Public review period: May-June
5. Final publication: end of June

ONTOLOGIES

Preparations for OWL and SHACL activity

- Objectives:
  - investigate current ontology activities around DCAT-AP
  - propose a common ontology
  - create W3C SHACL shapes for validation
- Approach:
  - small group of people with experience and expertise
  - drafts to be shared with the wider Working Group
- Planning
  - Start: end of March
  - Completion: mid June