DCAT-AP WEBINAR

25 OCTOBER 2016

v0.04

TOPICS OF TODAY

1. Opening
2. Issues that could be solved in additional guidelines
3. Issues that are change requests for a new revision of DCAT-AP
4. Issues that require liaison with others
5. Issues for SDSVoc
6. Next steps

OPENING | UPDATE

OPENING | TOUR DE TABLE

OPENING | OBJECTIVES OF THE WORK

- Follow up on the **discussion points** identified during the DCAT-AP workshop organised in May 2016.
- Collecting **new change requests** for the DCAT-AP (if any) via interactions with implementers through the DCAT-AP community channels;
- **Update existing DCAT-AP guidelines** to incorporate feedback; and
- Create a number of **new practical guidelines** that will support the Member States to implement DCAT-AP, focusing especially on the organisational and legal levels;

OPENING | OBJECTIVES OF THIS WEBINAR

- Verify with the group that the issues listed are relevant for guidelines
- Discuss possible further development of DCAT-AP
- Look at the wider issues that could/should be on the agenda of SDSVoc
- Plan the work ahead, especially in gathering legal and organisational issues

ISSUES FOR ADDITIONAL GUIDELINES | OVERVIEW

- Relationship between accessURL and downloadURL: practical use
- Publisher vs. contact point
- Contradiction dct:spatial and dct:Location
- De-referencing vocabularies
- URIs for organisations
- Entity-ID service
- Partitioning large DCAT-AP metadata files (in case of bulk harvesting)
- Agent roles

ISSUES FOR ADDITIONAL GUIDELINES

RELATIONSHIP BETWEEN ACCESSURL AND DOWNLOADURL: PRACTICAL USE

- dcat:accessURL (mandatory)

  *"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**."*

- dcat:downloadURL

  *"This property contains a URL that is a direct link to a **downloadable file** in a given format."*

- In some cases, accessURL might not be needed and the information in downloadURL and accessURL is duplicated. How to deal with this?

ISSUES FOR ADDITIONAL GUIDELINES

PUBLISHER VS. CONTACT POINT

- Publisher: the real-world entity of class foaf:Agent
- Contact point: the contact information of class vCard:Kind associated to an organisation, not the organisation itself
- The modelling difference may be hard to understand for implementers. Guidelines could be helpful in order to explain the difference.

ISSUES FOR ADDITIONAL GUIDELINES

CONTRADICTION DCT:SPATIAL AND DCT:LOCATION

- dct:spatial

  "The **MDR** Named Authority Lists must be used for continents, countries and places that are in those lists; if a particular location is not in one of the mentioned Named Authority Lists, **Geonames URIs** must be used."

- dct:location

  "A spatial region or named place. It can be represented using a controlled vocabulary or with geographic coordinates. In the latter case, the use of the **Core Location Vocabulary** is recommended, following the approach described in the GeoDCAT-AP specification."

- The use of the MDR vs. Geonames vs. the Core Location Vocabulary URIs seems to be confusing. A new guideline could clarify, e.g. by giving examples of actual use.

ISSUES FOR ADDITIONAL GUIDELINES

DE-REFERENCING VOCABULARIES

- Some of the controlled vocabularies specified for DCAT-AP are published on the **vocabulary level** in such a way that the **individual vocabulary terms** are **not de-referenceable**.
- E.g., it took some time for a dct:accessRights vocabulary to become available
- A **common approach** towards handling such situations, and a call for resolution of this problem will help implementers create metadata and support better interoperability.
- Guidelines could include examples of direct resolution of a term, access to a schema, and what to do if nothing comes back.

ISSUES FOR ADDITIONAL GUIDELINES

URIS FOR ORGANISATIONS

- DCAT-AP models publishers as **Agents**, which means that they should be identified by **URIs**. However, a lack of URIs for organisations has been reported.
- It might be useful to investigate the **existence** of sources for such URIs, like for example, the MDR Name Authority List for Corporate Bodies.
- Sharing practice on the **creation and maintenance** of such URI sources would be helpful for local implementations and for interoperability, e.g. examples of actual URIs, including 303 redirects and fragment identifiers (has URIs)

ISSUES FOR ADDITIONAL GUIDELINES

ENTITY-ID SERVICE

- The need to have a "third-party" (public maybe) system (provided as a service) that enable the maintenance (creation, manage, update, delete) of **persistent identifiers** (URIs) for
  - the entities mentioned in the DCAT profile (first of all, like persons and organizations) and
  - available to be used as well for other for entities mentioned in the published opendata datasets (inside the data)
- Support the **reuse of such identifiers** in the creation of several independent datasets;
- A public ID (URI) naming service independent from the specific context.
- Do you know any examples?

ISSUES FOR ADDITIONAL GUIDELINES

PARTITIONING LARGE DCAT-AP METADATA FILES (IN CASE OF BULK HARVESTING)

- In practical cases, DCAT-AP metadata is stored in a **single file** that contains the descriptions of the catalogue, all the datasets and all the distributions.
- For better management, it may be useful to share approaches to **split** such a single file into smaller files.
- A **common way** of doing this would enable development and deployment of **common tools** for metadata management and export.
- **Is this really a problem?**

ISSUES FOR ADDITIONAL GUIDELINES

AGENT ROLES: ADDITIONAL SET OF PROPERTIES LINKING TO AGENTS

Who is the right holder of a dataset?

It is possible to add multiple publishers to the dcat:Dataset, but currently there is no way to model that some publisher is a distributor/original publisher/rights holder. Possible solutions:

- Add separate distributor/publisher/rightsHolder **associations** from **Dataset to Agent**.
- Create **controlled vocabulary** that can be used to clarify the role of the Agent
- GeoDCAT-AP and national profiles could be used as a basis for the guideline

ISSUES FOR ADDITIONAL GUIDELINES

ANY OTHER ISSUES

ISSUES FOR REVISING DCAT-AP

MODELLING DATA QUALITY

- Data quality: precision, accuracy, fit-for-purpose, compliance with benchmarks, QA, etc.
- How to express data quality in DCAT-AP and its extensions?
- Several options exist: W3C Data Quality Vocabulary (DQV), dct:conformsTo, etc.
- It may be worth agreeing on a **consistent approach** to deal with different use cases: metadata conformance, data conformance, quality report, spatial resolution, quantitative QA results, user's feedback, etc.

ISSUES THAT REQUIRE LIAISON WITH OTHERS

- Mapping DCAT-AP to and from existing dataset publication platforms, e.g. CKAN, OpenDataSoft, Socrata, DKAN.
- Are you interested to **contribute** to those liaisons?

ISSUES FOR SDSVOC | OVERVIEW

- Relationships between Datasets, incl. versioning, time sequence, parent/child and grouping of collections: use of relation types
- Rights and licences for datasets: relationship with licences on catalogue and distributions
- Agent roles: additional set of properties linking to agents
- Service-based data access: modelling of non-file distributions and set of properties to enable machine-processing
- Relationship between Distributions: similarity criteria
- Packaging of distribution files: expression of format of included files
- Scientific data and data citation

ISSUES FOR SDSVOC | RELATIONSHIPS BETWEEN DATASETS

Evolutionary relation

***Versioning**. Each new version has newer data and replaces the precedent one, which turns obsolete and remains there for reference only*

Time series

***Sequence of points** in time observations*

Datasets with **parts**

***Parent** datasets with multiple **child** sub-datasets*

Collections

***Grouping** datasets under one umbrella, based on different dimensions, e.g. geography, or use cases*

ISSUES FOR SDSVOC | RIGHTS AND LICENCES

- In the DCAT model, rights and licences are assigned to **catalogues** and to distributions, not to datasets.
- In actual implementations, rights and licences may be associated with the **dataset**, applying to all distributions of the dataset.

ISSUES FOR SDSVOC | SERVICE BASED DATA ACCESS

- Many datasets are not published as files, but can be accessed through APIs or SPARQL endpoints
- Definition of Distribution in DCAT mentions that “Examples of distributions include a downloadable CSV file, an API or an RSS feed”.
- DCAT only seems to focus on **files**, for example by defining format and media type which are not relevant for APIs or end points.
- Specific information is necessary to access APIs and end points, e.g. methods Foad schemas, and the current version of DCAT does not include properties to express those types of information.

ISSUES FOR SDSVOC | RELATIONSHIPS FOR DISTRIBUTIONS

- DCAT definition of Distribution is ambiguous

  Represents a specific available form of a dataset. Each dataset might be available in different forms, these forms might represent different formats of the dataset or different endpoints. Examples of distributions include a downloadable CSV file, an API or an RSS feed
- Do all distributions contain the same data?
- May distributions contain different slices of a dataset?

  E.g. files for individual years in a multi-year dataset.
- Need to develop clear criteria to determine whether two data files or feeds can be distributions of a single dataset or of different datasets.

ISSUES FOR SDSVOC | PACKAGING OF DISTRIBUTION FILES

- Distributions are often made available in a packaged or compressed format
  E.g. a group of XLS files packaged in a ZIP file, compression of large files
- DCAT requires the package format to be expressed in dct:format
- It might be useful for an application to know what formats are contained in the package
- As a consequence, it might be useful if DCAT considered ways to indicate various levels of packaging. An example of an approach is in the way ADMS defines Representation Technique (see https://www.w3.org/TR/vocab-adms/#representation-technique).

ISSUES FOR SDSVOC | SCIENTIFIC DATA & DATA CITATION

In order to meet the requirements of the JRC Data Catalogue (http://data.jrc.ec.europa.eu/), which include support to data citation, JRC developed an extension of DCAT-AP for research data, covering the following requirements:

a) ability to indicate dataset **authors**

b) ability to describe data **lineage**

c) ability to give potential data consumers information on how to **use the data** ("usage notes")

d) ability to link to scientific **publications** about a dataset

e) ability to link to **input** data (i.e., data used to create a dataset)

Do you have similar requirements?

ISSUES FOR SDSVOC | MEETING DETAILS

- Clarify the steps needed to **improve communication** between data repositories and applications that use that data
- Potential outcome: a new W3C Working Group chartered to **extend DCAT** and determine how human and machine-readable metadata profiles are **defined** and made **discoverable**.
- Explore how W3C can best **support vocabulary development** for a variety of communities.

30 November - 1 December

Amsterdam, The Netherlands

For more info: https://www.w3.org/2016/11/sdsvoc

NEXT STEPS

- Today: first webinar
- 30/11/2016 – 01/12/2016: SDSVoc Amsterdam
- November 2016 – March 2017: interview rounds to identify legal and organisational issues

  Are you interested to participate?
- March 2017: second webinar: presentation of the new guidelines

Promoting semantic interoperability in Europe

STAY CONNECTED!

PROJECT OFFICERS

Vassilios.Peristeras@ec.europa.eu
Athanasios.Karalopoulos@ec.europa.eu

GET INVOLVED

- Follow @SEMICeu on Twitter
- Join the SEMIC goup on Linkedin
- Join the SEMIC community on Joinup

VISIT OUR INITIATIVES

- ADMS — ASSET DESCRIPTION METADATA SCHEMA
- StatDCAT-AP — FOR STATISTICAL DATASETS
- GeoDCAT-AP — FOR GEOSPATIAL DATASETS
- DCAT-AP — FOR DATA PORTALS IN EUROPE
- CORE PUBLIC ORGANISATION VOCABULARY
- CORE PERSON VOCABULARY
- REGISTERED ORGANISATION VOCABULARY
- CORE CRITERION & EVIDENCE VOCABULARY
- CORE LOCATION VOCABULARY
- CORE PUBLIC SERVICE VOCABULARY