SDSVoc Position statement

Makx Dekkers/2016-10-09/v0.02

1 Introduction

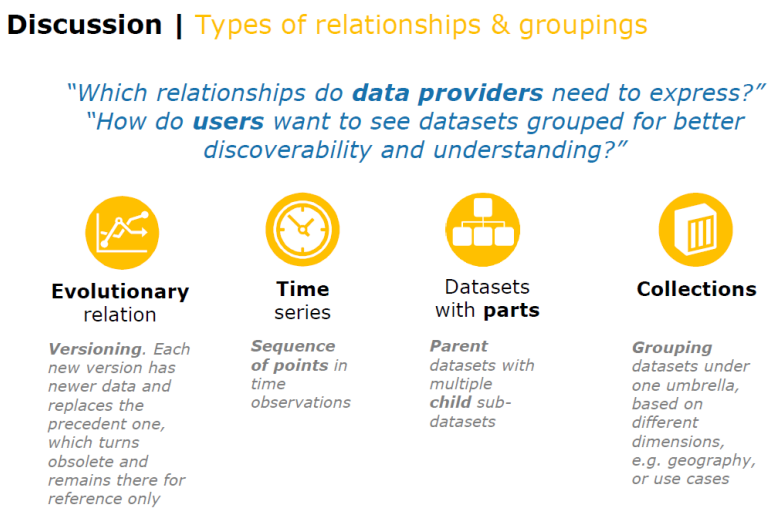
In work that I have done over a number of years, several issues have come up that may require further clarifications for the correct and interoperable use of DCAT and related recommendations.

2 DCAT issues

2.1 Relationships between datasets

In the specification of DCAT, datasets are treated as independent conceptual entities, only related to the catalogue of which they are part. However, in practical cases there may be several types of relationships between datasets for which there is no standard or recommended way to express them.

Several relationship types have been identified in the figure below that was used as a discussion slide during the DCAT-AP meeting on 13 May 2016 in Rome:



One of the DCAT-AP guidelines developed in 2015 (see https://joinup.ec.europa.eu/node/150348) suggests that providers focus on the expectations of the users and gives some possible approaches including the use of dct:hasPart and dct:hasVersion to handle some of these situations.

However, a fully interoperable approach might require additional properties and associated guidelines for DCAT. It would be useful if an analysis of actual requirements and practical approaches were to be conducted, leading to sharpened definitions and guidance with the possible addition of properties (e.g. sub-properties of dct:relation) to the DCAT Recommendation.

2.2 Distribution options

A large controversy emerged around the way that distributions of a single dataset may be related. The definition of a Distribution in DCAT 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*" as it does not make clear what *a specific available form* may contain.

Does it mean that all distributions contain the same data (e.g. the same observations), or may distributions contain different slices of the dataset, such as files for individual years in a multi-year dataset? The definition in DCAT is read by many to mean the first – the same set of observations in each of the distribution only differing in format – but there are some very strong opinions and arguments that favour the latter interpretation (e.g. see https://joinup.ec.europa.eu/asset/dcat_application_profile/issue/mo12-grouping-datasets#comment-16648).

In the current situation, a variety of approaches can be observed. In an analysis of the data in the DataHub (see https://joinup.ec.europa.eu/asset/dcat-ap_implementation_guidelines/issue/mi2-dataset-series#comment-17725) at least five different approaches could be observed.

Although it is probably too late to recommend a consistent approach given the existence of widely varying practices, it might be useful to develop clear criteria to determine whether two data files or feeds can be distributions of a single dataset or of different datasets – in which case the previous point comes into play, i.e. how to express the relationship between those datasets.

2.3 Non-file distribution and service-based data access

It turns out that many datasets in the wild are not published as files but can be accessed through services, APIs or SPARQL endpoints. The definition of Distribution in DCAT mentions that "*Examples of distributions include a downloadable CSV file, an API or an RSS feed*". However, 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 services, APIs and end points, e.g. methods ad schemas, and the current version of DCAT does not include properties to express those types of information.

It would be useful if DCAT were extended to take into account typical situations for common types of non-file distributions, identifying requirements for descriptive elements to support machine-processability.

A simple case is the one included in the specification of the StatDCAT-AP https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/asset_release/statdcat-ap-draft-4, where a property dct:type is added to the description of the Distribution with value http://publications.europa.eu/resource/authority/distribution-type/VISUALIZATION.

A proposal for the modelling of service-based access can be seen at https://joinup.ec.europa.eu/asset/dcat_application_profile/issue/dt2-service-based-data-access.

2.4 Packaged distributions

In practice, distributions are sometimes made available in a packaged or compressed format. For example, a group of files may be packaged in a ZIP file, or a single large file may be compressed. The current specification of DCAT allows the package format to be expressed in dct:format or dcat:mediaType but it is currently not possible to specify what types of files are contained in the package.

Therefore, it might be useful for DCAT to consider 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).

2.5 Datasets and catalogues

The DCAT model contains a hierarchy of the main entities: a catalogue contains datasets and a dataset has associated distributions. This model does not contemplate a situation that datasets exist outside of a catalogue, while in practice datasets may be exposed on the Web as individual entities without description of a catalogue.

Also, it may be inferred from the current model that a dataset, if it is defined as part of a catalogue, is part of only one catalogue; no consideration is given to the practice that datasets may be aggregated – for example when the European Data Portal aggregates datasets from national data portals.

It might be useful for DCAT to further clarify the relationships between datasets and zero, one or multiple catalogues. In particular, consideration of approaches to harvesting and aggregation – when descriptions of datasets are copied from one catalogue to another – contemplating the way that relationships between the descriptions can be maintained and how identifiers can be assigned that allow for linking back to the source descriptions.

3 Cross-vocabulary relationships

In the context of W3C working and interest groups (e.g. SWIG, GLD, DWBP) several overlapping vocabularies have been developed for the description of datasets: DCAT, VoID and Data Cube. These vocabularies define similar concepts, but it is not entirely clear how these concepts are related. For example, all three vocabularies define a notion of 'dataset' – dcat:Dataset, void:Dataset and qb:DataSet. These notions are similar but not entirely equivalent. For example, it has been argued that void:Dataset and qb:DataSet are more like a dcat:Distribution than a dcat:Dataset.

There is a need for clarification of how these approaches are similar or different and how they interact, for example in the form of guidelines how to create a DCAT description of a VOID or Data Cube dataset.