StatDCAT-AP: Representing statistical metadata by using the "DCAT application profile for data portals in Europe"

Marco Pellegrino
European Commission, Eurostat, marco.pellegrino@ec.europa.eu

Abstract: The StatDCAT Application Profile is an extension of the DCAT Application Profile for Data Portals in Europe: its purpose is to facilitate a better integration of the existing statistical data portals with the Open Data Portals, improving the visibility and discoverability of statistical datasets.
The final specification of version 1.0, released in December 2016, is currently being tested in a number of statistical data catalogues, one of them being the European Union's Open Data Portal run by the EU Publication Office.
In this context, the use of transformation mechanisms allows organisations using existing statistical standards for data and metadata exchange, such as SDMX, to align their standard with StatDCAT-AP in an easy manner.

1 Introduction

Open data portals have been established throughout Europe, in the last few years, interconnecting different types of data resources. As statistical data are of great interest for decision-making and for research purposes, the catalogues of open data portals also include numerous statistical datasets disseminated by statistical institutes and international organisations.

The StatDCAT Application Profile (StatDCAT-AP[1]) - an extension of the DCAT-AP[2] open data standard - supports the integration of descriptive metadata of statistical datasets within the catalogues of open data portals, hence improving the visibility and discoverability of statistical datasets.

StatDCAT-AP has been developed by a working group co-chaired by Eurostat and the EU Publication Office, within the framework of the ISA programme of the European Commission. The creation process of the new specification was open, transparent, and involved the main stakeholders to reach consensus in an open collaboration. This collaborative work applied and exploited the technical standards developed by W3C towards a globally interoperable environment of Linked Open Data. After a period of public review in summer-autumn 2016, StatDCAT-AP version 1 was published at the end of 2016.

[1] https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/home

[2] European Commission. ISA – DCAT Application Profile for data portals in Europe. https://joinup.ec.europa.eu/asset/dcat_application_profile/home

2 Background

The development of StatDCAT serves two main purposes:

a) Enhancing the interoperability between descriptions of statistical data sets and general data portals, facilitating the referencing of statistical data with other types of open data, thanks to a common RDF representation.

b) Facilitating a complete mapping between the RDF framework of standards and the SDMX Information Model. This mapping enables organisations using SDMX to know which metadata structures to use and how, in order to generate RDF-compliant messages directly from their SDMX metadata repositories. The co-existence of StatDCAT-AP with another DCAT profile for geospatial data (GeoDCAT-AP, see parallel paper on GeoDCAT by Perego et al.) is a further step towards a standard mapping between RDF, statistical standards and geospatial standards to improve interoperability and to integrate data production systems.

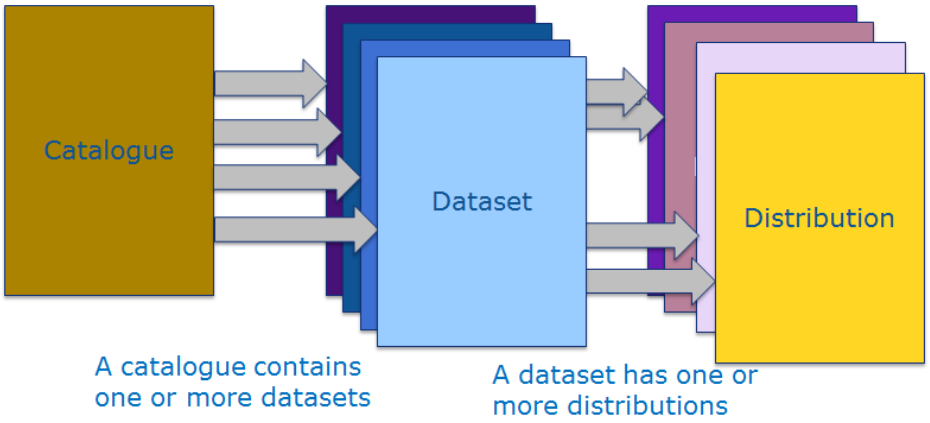
Building upon those two pillars, on one hand subscribing to the organisational goals to open up public data for cross-referencing, and on the other hand applying the emerging technologies that facilitate linking data together, StatDCAT-AP aims to improve the opportunities for discovery and reuse of statistical data from the wide audience using open data portals. In this context, the use of transformation mechanisms allows organisations using existing statistical standards for data and metadata exchange, such as SDMX, to align their standard with the StatDCAT-AP metadata specification.

3 What is StatDCAT-AP

StatDCAT-AP is fully aligned to DCAT-AP and to the W3C's Data Catalogue vocabulary (DCAT).

The original DCAT data model includes the following main entities:

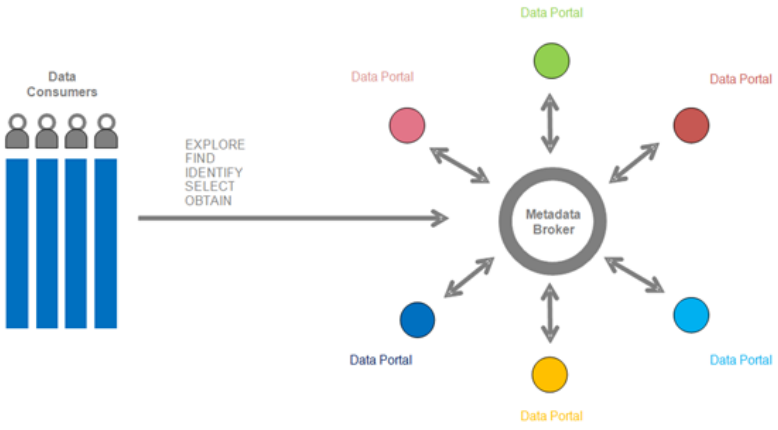
- The Catalogue, representing a collection of Datasets and defined as “a curated collection of metadata about datasets”.
- The Catalogue Record, which is “a record in a data catalogue, describing a single dataset”.
- The Dataset, representing the published information, and defined as “a collection of data, published or curated by a single agent, and available for access or download in one or more formats”.
- The Distribution, which represents a specific available form of a dataset. 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.



DCAT main entities

The basic use case of DCAT-AP is to make data better searchable across borders and sectors, by enabling a cross-data portal search for datasets.

There are two enabling conditions behind this "portal search". First, the data portals maintain a data catalogue including a collection of datasets, and making the descriptive metadata of the datasets freely available. Second, in order to maximise the interoperability, these descriptions should adhere to the DCAT-AP specifications for metadata. Thanks to these two conditions, a metadata *broker* can harvest catalogues of metadata from various data portals, delivering metadata in a validated and harmonised manner to data consumers. This creates the opportunity for professional communities to hook onto the emerging system of interoperable portals by aligning with the common metadata format.



DCAT-AP basic use case: enable a search for datasets across various data portals

QUICK REFERENCE OF DCAT CLASSES AND PROPERTIES

Class	Class URI	Mandatory properties	Recommended properties	Optional properties
Agent	foaf:Agent	foaf:name	dct:type	
Annotation	oa:Annotation			
Attribute Property	qb:AttributeProperty			
Catalogue	dcat:Catalog	dcat:dataset dct:description dct:publisher dct:title	foaf:homepage dct:language dct:license dct:issued dcat:themeTaxonomy dct:modified	dct:hasPart dct:isPartOf dcat:record dct:rights dct:spatial
Catalogue Record	dcat:CatalogRecord	dct:modified foaf:primaryTopic	dct:conformsTo adms:status dct:issued	dct:description dct:language dct:source dct:title
Category	skos:Concept	skos:prefLabel		
Category Scheme	skos:ConceptScheme	dct:title		
Checksum	spdx:Checksum	spdx:algorithm spdx:checksumValue		
Dataset	dcat:Dataset	dct:description dct:title	dcat:contactPoint dcat:distribution dcat:keyword dct:publisher dcat:theme	dct:accessRights dct:conformsTo foaf:page dct:accrualPeriodicity dct:hasVersion dct:identifier dct:isVersionOf dcat:landingPage dct:language adms:identifier dct:provenance dct:relation dct:issued adms:sample dct:source dct:spatial dct:temporal dct:type dct:modified owl:versionInfo adms:versionNotes
Dimension Property	qb:DimensionProperty			
Distribution	dcat:Distribution	dcat:accessURL	dct:description dct:format dct:license	dcat:byteSize spdx:checksum foaf:page dcat:downloadURL dct:language dct:conformsTo dcat:mediaType, subproperty of dct:format dct:issued

				dct:rights adms:status dct:title dct:modified
Document	foaf:Document			
Frequency	dct:Frequency			
Identifier	adms:Identifier	skos:notation		
Kind	vcard:Kind			
Licence Document	dct:LicenseDocument	dct:type		
Linguistic System	dct:LinguisticSystem			
Literal	rdfs:Literal			
Location	dct:Location			
Media Type or Extent	dct:MediaTypeOrExtent			
Period Of Time	dct:PeriodOfTime			schema:startDate schema:endDate
Provenance Statement	dct:ProvenanceStatement			
Publisher Type	skos:Concept			
Resource	rdfs:Resource			
Rights Statement	dct:RightsStatement			
Size or duration	dct:SizeOrDuration			
Standard	dct:Standard			
Status	skos:Concept			

StatDCAT-AP defines a small number of additions to the DCAT-AP model that are particularly relevant for statistical datasets: these are listed in the following table.

Given that there are many statistical datasets that are of interest to the general data portals and their users, it is likely that by recognising and exposing the additions to DCAT-AP proposed by StatDCAT-AP, general data portals will be able to provide enhanced services for collections of statistical data.

A revision of StatDCAT-AP to extend its scope to other metadata elements could be envisaged in the near future, as a consequence of the first pilot implementations. This could lead to make a distinction between a "core" and an "extended" StatDCAT-AP, where the second would cover more statistical metadata elements, for instance related to methodology and data quality.

STATDCAT-AP NEW PROPERTIES

Class URI	Type	Description
stat:attribute	Optional property (Dataset)	Range: qb:AttributeProperty Cardinality: 0..n This property links to a component used to qualify and interpret observed values, e.g. units of measure, any scaling factors and metadata such as the status of the observation (e.g. estimated, provisional). Attribute is a 'conceptual' entity that applies to all distribution formats, e.g. in case a dataset is provided both in SDMX and in Data Cube.
stat:dimension	Optional property (Dataset)	Range: qb:DimensionProperty Cardinality: 0..n This property links to a component that identifies observations, e.g. the time to which the observation applies, or a geographic region which the observation covers. Dimension is a 'conceptual' entity that applies to all distribution formats, e.g. in case a dataset is provided both in SDMX and in Data Cube.
stat:numSeries	Optional property (Dataset)	Range: rdfs:Literal typed as xsd:integer This property contains the number of data series contained in the Dataset. "Cartesian Product of the number of modalities of each dimension, excluding what Data Cube calls the measure dimension (that denotes which particular measure is being conveyed by the observation)". The numSeries is the actual number of series in the data set as referenced in the Distribution. This is usually less than the theoretical number calculated as the Cartesian Product (and sometimes significantly less). The actual number of series is, when combined with the dimension list, a useful indication of the detail of the data in the data set.
dqv:hasQualityAnnotation	Optional property (Dataset)	Range: oa:Annotation Cardinality: 0..n This property links to a statement related to quality of the Dataset, including rating, quality certificate, feedback that can be associated to the Dataset.
stat:statUnitMeasure	Optional property (Dataset)	Range: skos:Concept Cardinality: 0..n This property links to a unit of measurement of the observations in the dataset, for example Euro, square kilometre, purchasing power standard (PPS), full-time equivalent, percentage. Unit of measurement is a 'conceptual' entity that applies to all distribution formats, e.g. in the case when a dataset is provided both in SDMX and in Data Cube.
dct:type	Optional property (Distribution)	Range: rdfs:Resource Cardinality: 0..1 This property links to a type of the Distribution, e.g. that it is a visualisation.

4 Lessons learnt and future work

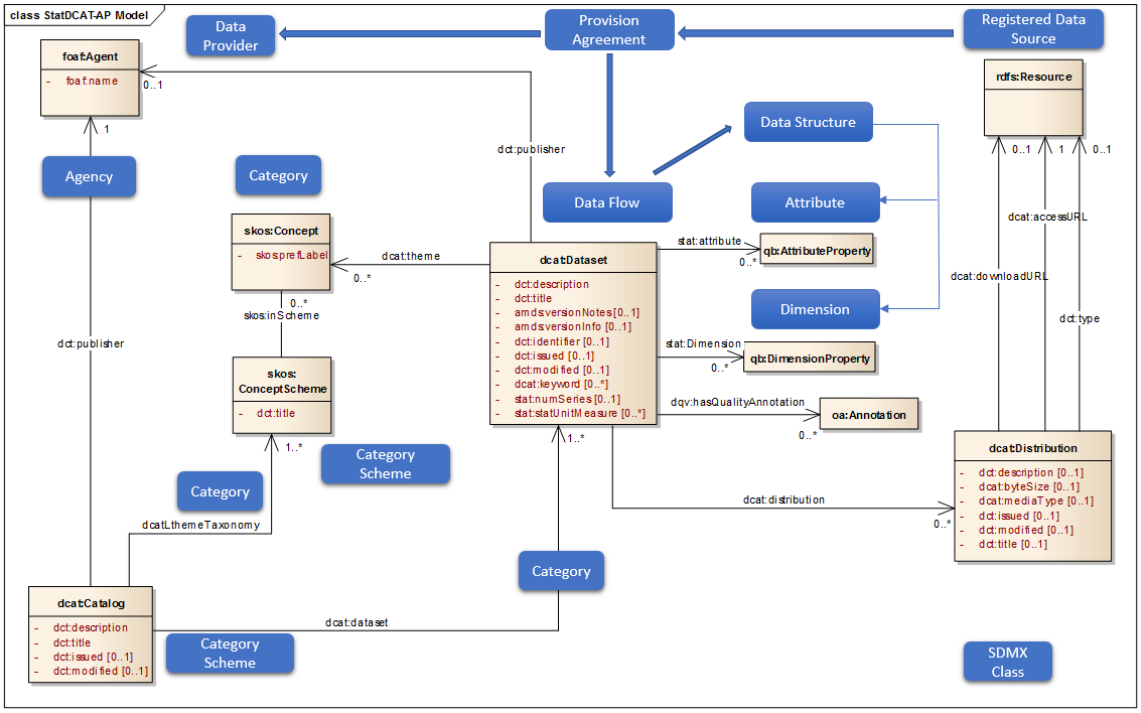
The development of StatDCAT-AP has demonstrated the importance of aligning metadata and its terminology to well-known and implemented frameworks, such as ISA Core Vocabularies, RDF Data Cube Vocabulary, Eurostat metadata vocabularies, and SDMX.

In the recent past, seven international organisations that are producing and coordinating the dissemination of statistical data, including Eurostat, defined and adopted the SDMX standard for data and metadata exchange, which is now an ISO

standard (IS-17369). By harmonising the metadata descriptions provided by SDMX (e.g. data structures, standard code lists, quality descriptions and methodology) and open data standards, both worlds get better connected, improving at the end the discoverability of statistical datasets.

StatDCAT is not yet supported by a set of tools, but some tests are being conducted to prove that a "dissemination chain" based on SDMX is also able to produce StatDCAT-AP metadata through a simple transformation.

The StatDCAT-AP specification also includes a section describing the mapping of StatDCAT-AP to the SDMX Information Model. This is achieved by means of schematic diagrams of the SDMX Information Model and through a worked example where the SDMX-ML content is mapped to the classes and properties of DCAT-AP. We actually expect more transformations to become available in the future, as the architecture of the StatDCAT-AP transformation mechanism could be easily used for DDI or CSV. Some examples and pilot implementations are expected to be produced in the near future.



StatDCAT-AP Model mapped to SDMX Model Classes

StatDCAT-AP has been brought to the attention of standards and statistical bodies, including the High Level Group for the Modernisation of Official Statistics, the SDMX

Secretariat and working groups, and several working groups operating within the European Statistical System.

The alignment with DCAT and DCAT-AP opens the way to further developments, for instance in the use of common vocabularies for data quality, or in the adoption of good practices in the use of global and uniform resource identifiers (URI). Future implementation activities will provide useful suggestions for improving the existing specifications and for enhancing the entire ecosystem of open data portals.

5 References

[1] European Commission. ISA – Interoperability Solutions for European Public Administrations. http://ec.europa.eu/isa/about-isa

[2] European Commission. ISA – DCAT Application Profile for data portals in Europe. https://joinup.ec.europa.eu/asset/dcat_application_profile/home

[3] StatDCAT-AP: https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/home

[4] SDMX: https://sdmx.org

[5] DIGICOM (European Statistical System's project for Digital communication, User analytics and Innovative products): http://ec.europa.eu/eurostat/web/ess/digicom

[6] EU Open Data Portal: http://data.europa.eu/euodp

[7] European Data Portal: https://www.europeandataportal.eu