StatDCAT-AP
A Common Layer for the Exchange of Statistical Metadata in Open Data Portals

Makx Dekkers, Stefanos Kotoglou, Chris Nelson, Marco Pellegrino, Norbert Hohn, Vassilios Peristeras

1 mail@makxdekkers.com
2 stefanos.kotoglou@be.pwc.com
3 chris.nelson@metadatatechnology.com
4 European Commission, Eurostat, Marco.Pellegrino@ec.europa.eu
5 European Union Publications Office, Norbert.Hohn@publications.europa.eu
6 European Commission, DG DIGIT/ISA2, Vassilios.Peristeras@ec.europa.eu

Abstract. The StatDCAT Application Profile is an extension of the DCAT Application Profile for Data Portals in Europe, version 1.1 (DCAT-AP). Its purpose is to provide a specification that is fully conformant with DCAT-AP version 1.1 as it meets all obligations of the DCAT-AP Conformance Statement. Its basic use case is to facilitate a better integration of the existing statistical data portals with the Open Data Portals, improving the discoverability of statistical datasets. StatDCAT-AP defines a small number of additions to the DCAT-AP model that are particularly relevant for statistical datasets. This will be beneficial for the general data portals, enabling enhanced services for the discovery of statistical data.

Keywords. Data catalogue, DCAT, linked open data, LOD, RDF, SDMX

1 Introduction

Statistical information is as an area of high value in the G8 Open Data Charter[1]. The G8 Open Data Charter and in its EU implementation refer to the importance of "high value datasets". Statistical data were identified as one of the thematic categories among those "in highest demand from re-users across the European Union".

With the design of StatDCAT-AP, the European Commission aims at strengthening the discoverability of statistical data across European data portals. In the European Union, the statistical data domain was one of the first areas providing transparent and open access to the public. In the context of the EU ODP (European Union Open Data Portal[2]), the Publications Office of the EU and Eurostat[3] collaborate on the regular

[1]G8. Open Data Charter and Technical Annex: Policy paper, 18 June 2013. Available online: https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex#action-2-release-of-high-value-data

[2]https://data.europa.eu/euodp

referencing of Eurostat's datasets into the portal. For that, a mapping exists from the Eurostat metadata into the EU ODP metadata representation. The Publications Office is in the transition process to align with the latest version of DCAT-AP.
As Eurostat is the largest contributor of statistical datasets to the European data portal, StatDCAT-AP is a joint initiative by Eurostat and the Publications Office within the framework of the ISA programme[4] run by the European Commission's Directorate-General Informatics (DIGIT).
Starting from DCAT-AP, specific communities can extend the basic Application Profile to support description elements specific for their particular data. One example is the GeoDCAT-AP that provides a connection between geospatial data described with metadata based on geospatial metadata standards on one hand and general Open Data portals on the other hand. The development of StatDCAT-AP is in line with that approach by extending the DCAT-AP with descriptive elements that can further help in the discovery and use of statistical data sets.

2 Potential links with international initiatives on Linked Open Data

StatDCAT-AP is developed under the ISA programme of the EU. ISA (Interoperable Solutions for European Public administrations)and its follow-up ISA^{2}aim at ensuring interoperability between different IT systems responsible for the delivery of electronic services, thereby ensuring seamless electronic cross-border or cross-sector interaction between public administrations, businesses and citizens.

The approach used in StatDCAT also fits well in the general framework of two other international initiatives in the area of linked open data and could potentially serve for the description of use cases. These two initiatives are:

- The "Implementing Modernstats Standards" initiative which is the subject of a major international collaboration under the UNECE[5] High-Level Group for the Modernisation of Official Statistics[6]
- The DIGICOM project[7]of the European Commission

One of the deliverables of the "Implementing Modernstats Standards" project is to provide a central repository of key standards in the form of "linked open metadata", thus reducing the need for statistical organizations to develop such a resource individually (Work Package1: Build a dissemination system for core structural metadata).

[3]http://ec.europa.eu/eurostat
[4] http://ec.europa.eu/isa/isa2/index_en.htm
[5] United Nations Economic Commission for Europe.
[6]The mission of the UNECE High-Level Group for the Modernisation of Official Statistics (HLG-MOS) is to oversee the development of frameworks, tools and methods, to support modernisation in statistical organizations. The aim is to improve the efficiency of statistical production, and the ability to produce outputs that better meet user needs.
[7]DIGICOM stands for Digital communication, User analytics and Innovative products.

The DIGICOM project's goal is to create new, innovative dissemination products, tools and services for ESS[8] statistics. The idea is that Eurostat and National Statistical Institutes across Europe work together to develop solutions in four fields: a) innovative user interaction; b) modern visualization tools; c) easy access to data, linked open data; d) communication and promotion.

3 What is DCAT-AP - brief overview

The DCAT Application profile for data portals in Europe (DCAT-AP) is a specification based on W3C's Data Catalogue vocabulary (DCAT) for describing public sector datasets in Europe. Its basic use case is to enable a cross-data portal search for data sets and make public sector data better searchable across borders and sectors. This can be achieved by the exchange of descriptions of data sets among data portals.

The specification of the DCAT-AP was a joint initiative of DG CONNECT, the EU Publications Office and the ISA Programme. The specification was elaborated by a multi-disciplinary Working Group with representatives from 16 European Member States, some European Institutions and the US.

The first version (1.0) of the Application Profile was published in September 2013. In 2015, a revised version (1.1) was developed and published in November 2015 with changes based on requests from implementers of the first version.

The DCAT-AP data model includes the following main entities:

- The Catalogue: this represents a collection of Datasets. It is defined in the DCAT Recommendation[9] as "a curated collection of metadata about datasets". The description of the Catalogue includes links to the metadata for each of the Datasets that are in the Catalogue.
- The Catalogue Record: DCAT defines this as "a record in a data catalog, describing a single dataset". The Catalogue Record enables statements about the description of a Dataset rather than about the Dataset itself. Catalogue Records may not be used by all implementations. It is optional in DCAT-AP and mostly used by aggregators to keep track of harvesting history.
- The Dataset: this represents the published information. It is 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 description of a Dataset includes links to each of its Distributions, if they are available. A Dataset is not required to have a Distribution; examples are Datasets that are described before the associated data is collected, Datasets for which the data has been removed, and Datasets that are only accessible through a landing page.

[8]European Statistical System, i.e. the 28 Member States of the European Union, EFTA countries and EU candidate countries. http://ec.europa.eu/eurostat/web/ess/

[9]W3C. Data Catalogue Vocabulary. W3C Recommendation 16 January 2014. https://www.w3.org/TR/vocab-dcat/

- The Distribution: this, according to DCAT, "represents a specific available form of a dataset. Each dataset might be available in different forms, and these forms might represent different formats of the dataset or different end-points. Examples of distributions include a downloadable CSV file, an API or an RSS feed". The description of a Distribution contains information about the location of the data files or access point and about the file format and licence for use or reuse. In the case of statistical datasets, Distributions may be available in specific formats like SDMX-ML or using the Data Cube vocabulary.

The data model of version 1.1 of DCAT-AP is available at https://joinup.ec.europa.eu/system/files/project/dcat-ap_0.bmp.

The full version of the profile is posted on Joinup, the collaborative platform of the European Commission funded by ISA Programme[10].

3.1 The basic use case for DCAT-AP

The basic use case that this specification intends to enable is a cross-data portal search for datasets. This can be achieved by the exchange of descriptions of datasets among data portals. The basic use case involves the following actors and systems:

- Data providers: Data providers include a description of their datasets on one or more data portals, so that the datasets can be more easily found.
- Data portals: Data portals maintain a data catalogue including a collection of datasets made available by data publishers. Data portals make the description metadata of the datasets in their collection freely available to third parties. In addition, data portals may also make collections of relevant datasets of other data portals searchable via their user interface. For enhanced interoperability, the description metadata adheres to the specifications of the DCAT Application Profile.
- Metadata Brokers: Metadata Brokers facilitate the exchange of description metadata between data portals by ensuring conformance to the DCAT Application Profile. They provide metadata harvesting, transformation, validation, harmonisation, and publication services. The Open Data Support[11] project funded by the European Commission will operate a Metadata Broker service for data portals in Europe, it will use the DCAT Application Profile as a common metadata vocabulary
- Data Consumers: Users (data consumers) use the data portal of their choice to search through various collections of datasets from a single point of access. The data portal allows the user to explore (FRSAD – Functional Requirements for Subject Authority Data[12]), find, identify and select (FRBR –

[10]https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11.

[11] Open Data Support: https://joinup.ec.europa.eu/node/62928

[12]IFLA. Functional Requirements for Subject Authority Data (FRSAD). http://www.ifla.org/en/node/1297

Functional Requirements for Bibliographic Records[13]) datasets coming from different EU Member States, different portals and different organisations. Data consumers could also be systems (machines).

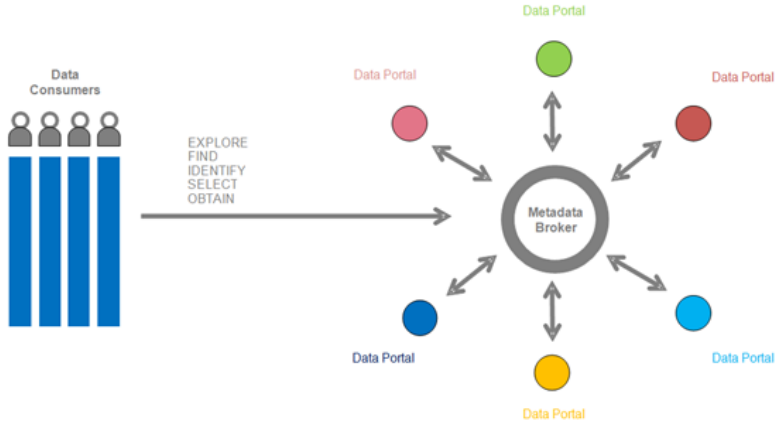

Figure 1 - Basic use case: enable a search for datasets across various data portals

3.2 Terminology used in the DCAT Application Profile

An Application Profile is a specification that re-uses terms from one or more base standards, adding more specificity by identifying mandatory, recommended and optional elements to be used for a particular application, as well as recommendations for controlled vocabularies to be used.

A Dataset is a collection of data, published or curated by a single source, and available for access or download in one or more formats.

A Data Portal is a Web-based system that contains a data catalogue with descriptions of datasets and provides services enabling discovery and re-use of the datasets.

In the following sections, classes and properties are grouped under headings 'mandatory', 'recommended' and 'optional'. These terms have the following meaning.

- Mandatory class: a receiver of data MUST be able to process information about instances of the class; a sender of data MUST provide information about instances of the class.
- Recommended class: a sender of data SHOULD provide information about instances of the class; a sender of data MUST provide information about instances of the class, if such information is available; a receiver of data MUST be able to process information about instances of the class.

[13]IFLA. Functional Requirements for Bibliographic Records. http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

- Optional class: a receiver MUST be able to process information about instances of the class; a sender MAY provide the information but is not obliged to do so.
- Mandatory property: a receiver MUST be able to process the information for that property; a sender MUST provide the information for that property.
- Recommended property: a receiver MUST be able to process the information for that property; a sender SHOULD provide the information for that property if it is available.
- Optional property: a receiver MUST be able to process the information for that property; a sender MAY provide the information for that property but is not obliged to do so.

The meaning of the terms MUST, MUST NOT, SHOULD and MAY in this section and in the following sections are as defined in RFC 2119[14].

4 What is StatDCAT-AP

DCAT-AP is intended as a common layer for the exchange of metadata for a wide range of dataset types. The availability of such a common layer creates the opportunity for professional communities to hook onto the emerging landscape of interoperable portals by aligning with the common exchange format. In addition to the basic DCAT-AP, specific communities can extend the Application Profile to support description elements specific for their particular data.

The development of a DCAT-AP extension for the exchange of metadata for statistical datasets is in line with that approach. The StatDCAT Application Profile is an extension of the DCAT Application Profile for Data Portals in Europe, version 1.1 (DCAT-AP). Its purpose is to provide a specification that is fully conformant with DCAT-AP version 1.1 as it meets all obligations of the DCAT-AP Conformance Statement. Its basic use case is to facilitate a better integration of the existing statistical data portals with the Open Data Portals, improving the discoverability of statistical datasets.

As a result, data portals that comply with DCAT-AP will be able to understand the core of StatDCAT-AP. In addition, StatDCAT-AP defines a small number of additions to the DCAT-AP model that are particularly relevant for statistical datasets. Given that there are many statistical datasets that are of interest to the general data portals and their users, it is likely that recognising and exposing the additions to DCAT-AP proposed by StatDCAT-AP will be beneficial for the general data portals to be able to provide enhanced services for collections of statistical data.

The work on StatDCAT-AP is a first activity in the context of a wider roadmap of activities that aim to deliver specifications and tools that enhance interoperability between descriptions of statistical data sets within the statistical domain and between statistical data and open data portals.

[14]IETF. RFC 2119. Key words for use in RFCs to Indicate Requirement Levels. http://www.ietf.org/rfc/rfc2119.txt

4.1 What's new in StatDCAT-AP?

The StatDCAT-AP data model includes the four main entities that are also present in DCAT-AP: Catalogue, Catalogue Record, Dataset and Distribution.

Discussions during the development of the StatDCAT-AP specifications brought out a number of requirements for the description of statistical datasets that were not met by existing properties in DCAT-AP. The following sections present the extensions that are included in StatDCAT-AP to meet those requirements.

4.1.1 Dimensions and attributes

A requirement has been expressed to expose information about:

- Attributes: components used to qualify and interpret observed values such as units of measure, scaling factors.
- Dimensions: components that identify observations such as time, sex, age, regions.

The following properties are created in the StatDCAT-AP namespace:

- stat:attribute: This property is to be used to identify an attribute that is used in the Dataset. Attributes enable specification of the units of measure, any scaling factors and metadata such as the status of the observation (e.g. estimated, provisional). The range of this property is: qb:AttributeProperty, expressed as a URI.
- stat:dimension: This property is to be used to identify a dimension that is used in the Dataset. Examples of dimensions include the time to which the observation applies, or a geographic region which the observation covers. The range of this property is: qb:DimensionProperty, expressed as a URI. The stat:dimensionreferences the dimension, not the terms that are in that dimension. This could indeed allow a richer discovery but would require a different property. The issue will be considered for a future update.

Another property is the stat:statMeasure, i.e. the unit of measurement of the observations in the dataset, e.g. Euro, square kilometer, percentage, full-time equivalent. The range of this property is: skos:Concept.

4.1.2 Quality aspects

Quality aspects are very important for datasets in general and for statistical datasets in particular. The current specification includes a first approach that allows certain annotations related to quality to be made, either through linking to quality information that is already published elsewhere or by including text with quality information.

The following annotation property is included, re-used from the Data Quality Vocabulary[15] that is being developed by the Data on the Web Working Group at W3C:

- dqv:hasQualityAnnotation: A statement related to quality of the Dataset, including rating, quality certificate, feedback that can be associated to datasets or distributions. The range of this annotation is: oa:Annotation.

A more structured solution would require more time and resources, and this was not possible for the current release. The description of quality aspects will be improved in the next update, taking into account existing experiences, other evolving standards and responding to any feedback deriving from the review and first experiences with version 1.0.

4.1.3 Visualisation

One of the requirements raised during discussions was the need to be able to link to a visualisation of the data, for example a document or Webpage where a tabular or graphical representation of the data can be viewed, or an interactive service where the data can be accessed and viewed.

The agreed approach for these types of visualisations is to model them as Distributions with a type of "visualisation" from the MDR Distribution type Named Authority List[16].

In order to implement this approach, a property to indicate the type of distribution is added to the class Distribution, re-used from Dublin Core:

- http://purl.org/dc/type: This property is the nature or genre of the resource. The property is to be used to indicate the type of a Distribution, in particular when the Distribution is a visualisation. The range of this property is: URI of a term in a controlled vocabulary.

Further use of the Type property on Distribution may be considered in the future, for example to indicate that data can be accessed through a service.

4.1.4 Other extensions

A requirement was brought forward to allow expression of the number of data series contained in a dataset. A series is a unique cross product of values of dimensions (excluding time). The number of data series therefore gives an indication of the potential scope of a data set. A Dataset could contain data for three regions with three values for each region. In this example, the number of series is three while the number of observations is nine.

[15] Data on the Web Best Practices: Data Quality Vocabulary. W3C Working Draft 19 May 2016. https://www.w3.org/TR/vocab-dqv/

[16] MDR Authorities. Distribution types. http://publications.europa.eu/mdr/authority/distribution-type/

This property is created in the StatDCAT-AP namespace:

- http://data.europa.eu/(xyz)/statdcat-ap/numSeries: The number of data series in a Dataset. The range of this property is: dct:SizeOrDuration, expressed as xsd:integer.

The UML diagram of StatDCAT-AP is presented in Figure 3.

5 How to produce StatDCAT metadata: an example based on SDMX

StatDCAT-AP focuses on metadata elements that may contribute to data discovery, encouraging the use of common controlled vocabularies and the re-use of metadata from existing repositories.

In the recent past, seven international organisations that are producing and coordinating the dissemination and sharing of statistical data, including Eurostat, defined and adopted the SDMX[17] standard for data and metadata exchange, which is now an ISO standard (IS-17369).

SDMX – recommended at United Nations level since 2008 and widely implemented by the European Statistical System and the European System of Central Banks – ensures that the exchange of statistical data happens without loss of information and following a clear information model.

By correlating 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.

The StatDCAT-AP specification 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.

[17] https://sdmx.org/

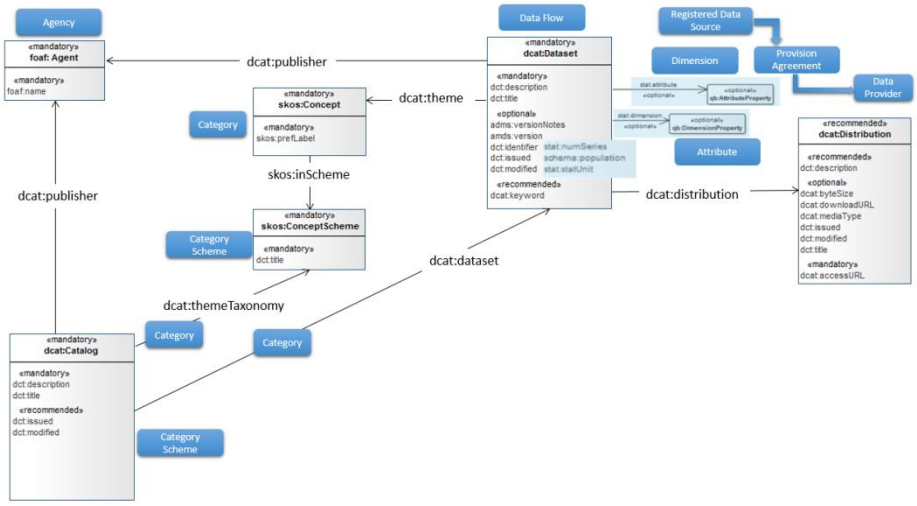

Figure 2: DCAT-AP Model mapped to SDMX Model Classes

The intent of this mapping is twofold:

1. It enables those organisations that are using SDMX to know which metadata structures to use in order to create StatDCAT-AP directly from existing SDMX metadata repositories (such as an SDMX Registry).
2. It enables organisations that wish to use SDMX structural metadata as the format for a Transformation Mechanism to know which SDMX element or attribute maps to which StatDCAT-AP class or property.

A dissemination chain based on SDMX data descriptions is also able to produce StatDCAT-AP descriptions through a simple transformation.

The StatDCAP-AP specification contains more technical documentation about these aspects, which are relevant for organisations using SDMX infrastructures. SDMX is one of the main standards currently in use in the statistics field (and this explains why we dedicated so much space to this point) but we actually expect more transformations to become available in the future, as the architecture of the StatDCAT-AP transformation mechanism lends itself just as easily to CSV and LOD as is does to SDMX. Some examples and pilot implementations are deemed to be produced and documented in the near future.

6 Conclusion

The StatDCAT-AP work is being conducted in a transparent manner, visible to the public, with the objective of involving the main stakeholders to reach consensus in an open collaboration. This collaborative work takes place in a wider context, both on the European level with the Directive on the re-use of Public Sector Information, and

on the global level with the G8 Open Data Charter. At the same time, it applies the technical standards developed by W3C towards a globally interoperable environment of Linked Open Data.

Building upon these two pillars, on one hand subscribing to the organisational goals to open up public data for reuse, 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 to a wide audience.

StatDCAT-AP creates the opportunity for statistical institutes and other professional communities to hook onto the emerging landscape of interoperable open data portals by aligning with the common exchange format. In this context, the use of transformation mechanisms such as the ones described for SDMX would make the job of implementing StatDCAT-AP much easier.

StatDCAT-AP is currently in a period of public review. The draft version 1 is available at https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/home.

The approval of version 1 for publication is planned for autumn 2016, after consolidation of any feedback received.

In the near future, we will set out to demonstrate the value and opportunities offered by StatDCAT-AP in practice and may report on the results of implementation at future events.

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. http://ec.europa.eu/isa/ready-to-use-solutions/dcat-ap_en.htm
3. StatDCAT-AP draft: https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/home

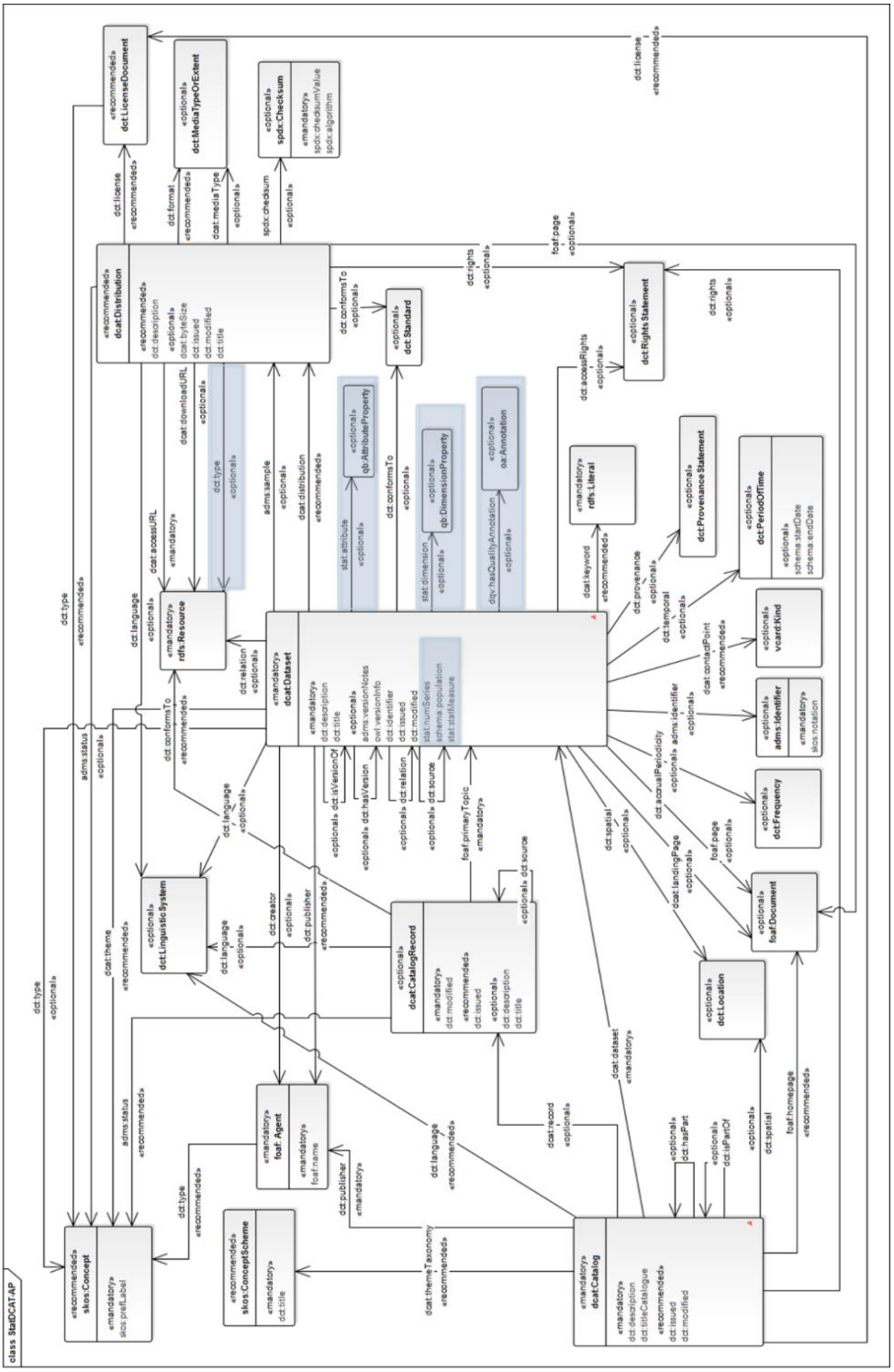

Figure 3: StatDCAT-AP Data Model