SC8DI07171

D02.01.01.02: StatDCAT-AP – DCAT Application Profile for description of statistical datasets, Draft 1

Document Metadata

Date	2016-03-08
Status	Internal draft
Version	0.04
Authors	Makx Dekkers – AMI Consult
Reviewed by	Nikolaos Loutas – PwC EU Services Marco Pellegrino – Eurostat Norbert Hohn – Publications Office
Approved by	

This report was prepared for the ISA Programme by:

PwC EU Services

Disclaimer:

Table of Contents

List of Tables

No table of figures entries found.

List of Figures

1 INTRODUCTION

1.1 Background

Collecting, compiling, analysing and publishing statistical data is a long standing method to support decision making. Statistical data is available via high-end quality data publishing platforms as well as as ad-hoc created tabular data. It has to be noted that the statistical data domain was one of the first data domains that was providing open and transparent access to its data.

This value has been recognised: statistical information has been identified as "high value datasets" in the G8 Open Data Charter[1] and in its EU implementation[2]. This statement is confirmed in the Commission's Notice 2014/C 240/01[3], elaborating the results of the online consultation launched by the Commission in August 2013 on the revision to the PSI Directive[4]. According to the feedback received, statistical data was identified as one of the thematic dataset categories among those "in highest demand from re-users across the EU".

At the same time, Open Data Portals are being established throughout Europe by EU Member States. On the European level, the European Data Portal[5] has started operation in November 2015. Statistical data is of great interest for all of the data categories in such open data portals and therefore it is beneficial for references to statistical datasets to be prominently visible in such data portals.

Open data portals bring together metadata, descriptions of datasets that are hosted by data providers. The portals harvest the metadata that is provided by the providers from their content management systems in a standard exchange format. This standard metadata exchange format is known as the DCAT Application Profile for data portals in Europe (DCAT-AP)[6], developed under the aegis of the European Commission's ISA programme[7].

Through 2015, activities have already taken place towards the scoping of the work on StatDCAT-AP. Preliminary work was done by a Core Working Group with representation from Eurostat, Publications Office, DG CONNECT and representatives of ISA supported by the contractor's experts. That earlier work included definition of some terminology (data vs. metadata), an analysis of the statistical data publishing field and an analysis of standards for publishing statistical data and metadata. A conceptual mapping of

[1] Gov.uk. Cabinet Office. G8 Open Data Charter and Technical Annex. Policy paper, 18 June 2013. Action 2: Release of high value data. https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex#action-2-release-of-high-value-data
[2] European Commission. Digital Agenda for Europe. EU Implementation of the G8 Open Data Charter. 31 October 2013. http://ec.europa.eu/digital-agenda/en/news/eu-implementation-g8-open-data-charter
[3] EUR-Lex. Commission notice — Guidelines on recommended standard licences, datasets and charging for the reuse of documents. OJ C 240, 24.7.2014, p. 1–10. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52014XC0724(01)
[4] EUR-Lex. Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information. http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1449913281728&uri=CELEX:32013L0037
[5] European Commission. European Data Portal. http://www.europeandataportal.eu/
[6] European Commission. ISA – Interoperability Solutions for European Public Administrations. DCAT Application Profile for data portals in Europe. http://ec.europa.eu/isa/ready-to-use-solutions/dcat-ap_en.htm
[7] European Commission. ISA – Interoperability Solutions for European Public Administrations. http://ec.europa.eu/isa/about-isa/

SDMX to DCAT-AP was also undertaken both on the metadata level (assessing "reference" metadata created by the ESMS as the source for creation of data set descriptions) and on the data level (assessing how "structural" metadata can be derived from the data structure definition). In addition, the metadata properties used in statistical data portals such as Eurostat were evaluated.

The final report[8] of the work done in 2015 is available from the European Commission's ISA programme.

1.2 Objectives

The 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 a wide range of 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 basic 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, called StatDCAT-AP, is in line with that approach, first by **determining which description elements in statistical data standards can be exposed in the DCAT-AP format**, and second by **extending the DCAT-AP with descriptive elements that can further help in the discovery and use of statistical data sets**.

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. This roadmap, outlined in the next section, includes several activities that take place over a longer period.

The work on the specification of the StatDCAT-AP contained in this document took place over a period of eight months from November 2015 through June 2016 and covered a set of initial activities in this context. The aim of this first step was that, within the available time and resources, concrete results could be achieved that act as a demonstration and a reality check for the roadmap.

The overall objective of this first phase of work is summarised in the following charter:

> The StatDCAT-AP activity is a first step in a roadmap that aims to enhance interoperability between descriptions of statistical data sets and general data portals, facilitating referencing of statistical data with other open data.
>
> The concrete objective of the work is to develop and reach consensus on an Application Profile of the Data Catalog Vocabulary (DCAT) to be used for the description of statistical data sets with an initial focus on discovery of those data sets in a wider context.
>
> The StatDCAT-AP will be based on the DCAT Application Profile for Data Portals in Europe (DCAT-AP). In addition, initial guidelines on the extraction of relevant metadata from the existing implementation at Eurostat and

[8] D02.01.2 Specification of StatDCAT-AP. A statistical extension for the DCAT application profile for data portals in Europe. Version 0.11. 2015-09-25. Available on request.

possibly others will be elaborated in order to enable the export of metadata conforming to the application profile from existing data.

Based on the contributions of the main stakeholders, extensions to DCAT-AP can be proposed with descriptive elements particularly useful for discovery of statistical data sets beyond the possibilities offered by the generic DCAT-AP.

The work in this phase will concentrate on use cases that improve the discovery of statistical data sets published in open data portals across European institutions and EU Member States and in particular in the European Open Data Portal, as well as use cases that facilitate the integration of statistical data sets with open data from other domains.

The participants in this work had the opportunity to collaborate with colleagues from the statistical domain and with experts from the open data community, contributing and sharing their knowledge and experience with the current implementations of the statistical data standards, and were able to gain insight into possible approaches by which statistical data can be better disclosed outside of the statistical domain.

1.3 Roadmap

The wider roadmap involves several steps as listed here:

1. Connecting descriptions of statistical datasets with general open data portals through a common basic exchange format, i.e. the StatDCAT-AP;
2. Developing guidelines for the extraction of metadata from specific implementations of statistical standards towards the common exchange format;
3. Harmonising implementations of statistical standards towards a more coherent landscape of statistical resources, possibly as an extension of the basic StatDCAT profile (for the metadata level) and through the use of W3C RDF Data Cube Vocabulary (for the data level),
4. Creating a set of tools to facilitate automatic extraction and validation of metadata from data described by statistical standards into StatDCAT-AP;
5. Conducting practical pilots in various stages of the above activities to test and verify approaches and solutions.

The work reported in this document covers the first two points of the roadmap.

1.4 Structure of this document

Outline of chapters.

2 TERMINOLOGY USED IN THIS DOCUMENT

All specific terminology, using as a main source the 2015 work on StatDCAT-AP, and further augmented through resolution of terminology issues raised in the Working Group.

3 RELATED WORK

3.1 Statistical data and metadata initiatives

3.1.1 Eurostat and EU Publications Office collaboration

In the context of the European Union Open Data Portal (EU ODP)[9], the Publications Office and Eurostat collaborate on the automated ingestion of Eurostat's datasets into the EU ODP. For that, there exists a mapping from the Eurostat metadata into the EU ODP metadata representation (a preliminary version of DCAT-AP)[10]. Today the Publications Office is in the transition process to align with DCAT-AP. As Eurostat is the largest contributor of datasets to EU ODP, StatDCAT-AP is a joint initiative by Eurostat and Publications Office to make more high quality metadata associated with the statistical datasets also available in a more general context of Open Data Portals.

The work is supported also by DG CONNECT, since the pan-European data portal will be one of the key implementers of the StatDCAT-AP as the common metadata standard for harmonising the descriptions of statistical datasets originating from different countries.

The Interoperability Solutions of European Public Administrations (ISA) Programme of the European Commission is, through ISA Action 1.1, the sponsor of the activity.

3.1.2 SDMX

SDMX[11], which stands for Statistical Data and Metadata eXchange is an international initiative that aims at standardising and modernising ("industrialising") the mechanisms and processes for the exchange of statistical data and metadata among international organisations and their member countries.

SDMX is sponsored by seven international organisations including the Bank for International Settlements (BIS), the European Central Bank (ECB), Eurostat (Statistical Office of the European Union), the International Monetary Fund (IMF), the Organisation for Economic Cooperation and Development (OECD), the United Nations Statistical Division (UNSD), and the World Bank.

These organisations are the main players at world and regional levels in the collection of official statistics in a large variety of domains (agriculture statistics, economic and financial statistics, social statistics, environment statistics etc.).

The main components of SDMX are presented in Figure 1.

[9] European Union Open Data Portal. http://open-data.europa.eu
[10] See the file ESTAT_xxx.zip in http://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing
[11] Statistical Data and Metadata eXchange. https://sdmx.org/

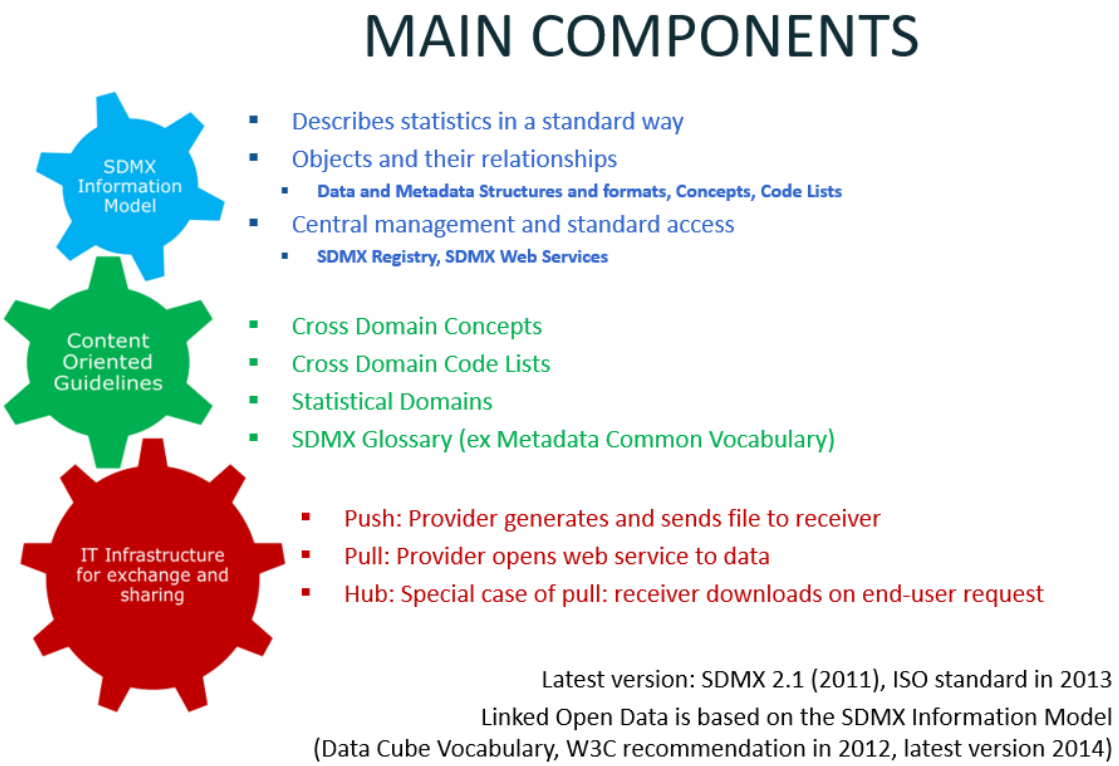

Figure 1: SDMX Main Components

A schematic view of the information model can be seen in Figure 2.

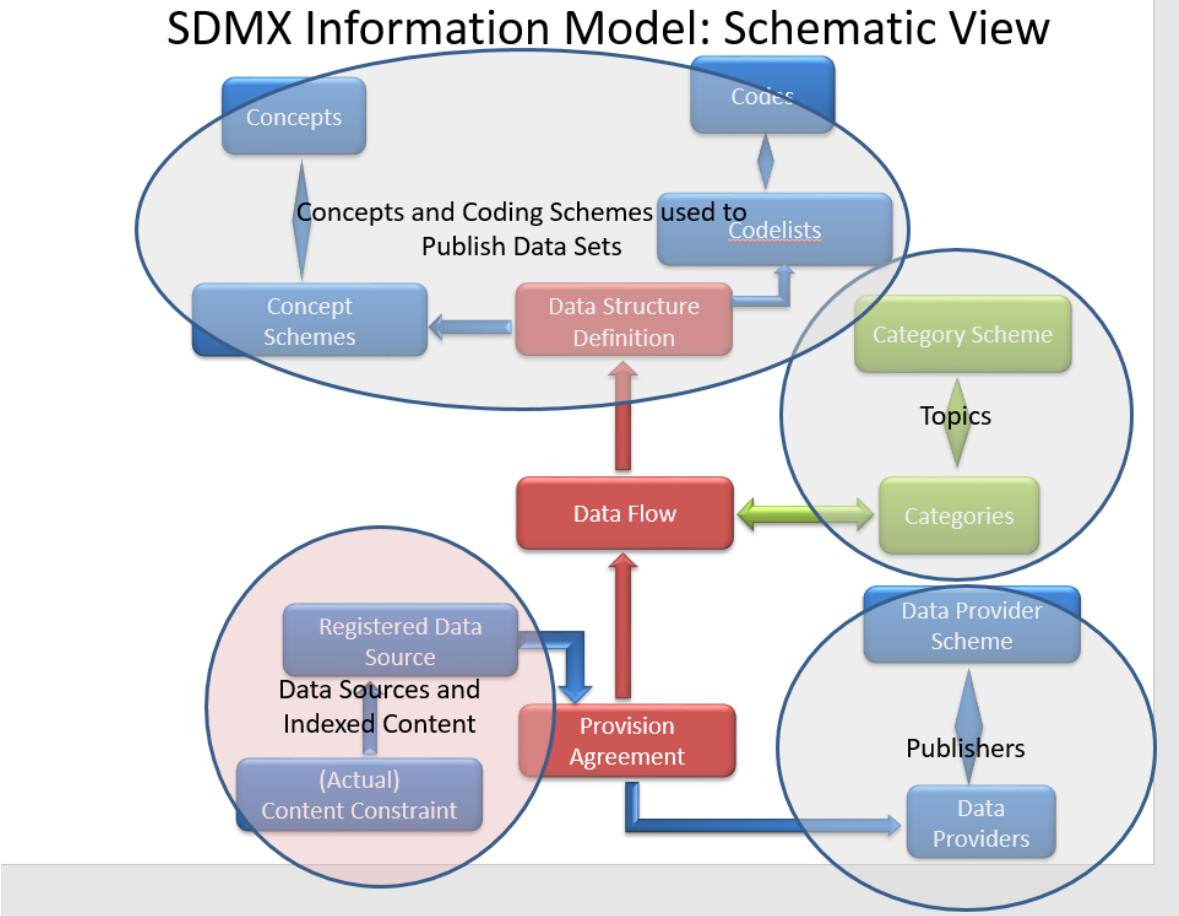

Figure 2: SDMX Information Model: Schematic View

3.1.3 ESMS

The Euro SDMX Metadata Structure (ESMS)[12] contains the description and representation of statistical metadata concepts to be used for documenting statistical data and for providing summary information useful for assessing data quality and the production process in general. The broad concepts used are compatible with the SDMX cross-domain concepts and with the common terminology as published within the SDMX "Metadata Common Vocabulary" (all published in January 2009).

The ESMS is addressed to the European Statistical System. It is implemented at Eurostat and at national level: the application of the concepts and sub concepts at European level and at national level is stated in the ESS guidelines.

The information to be entered is normally free text. Only in some cases, code lists will be used in the future: this is already indicated in the column "representation".

The ESMS allows the creation of different output files comprising information related to all the concepts listed or a subset of those concepts. These output files can be used for different purposes (data dissemination, quality reporting, etc.).

[12] Eurostat. Euro-SDMX Metadata Structure (ESMS). http://ec.europa.eu/eurostat/data/metadata

3.1.4 Others, e.g. OECD, World Bank, US…

Overview …

3.2 Open Data standards and application profiles

3.2.1 W3C DCAT

The basis for DCAT-AP is the specification of the Data Catalog Vocabulary (DCAT)[13]. DCAT was developed in the period from June 2011 through December 2013 by the Government Linked Data Working Group[14]. The specification was published as a W3C Recommendation in January 2014.

The abstract in the specification describes it as follows:

DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use.

By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation.[15]

The specification defines RDF Classes and Properties in a model that has four main entities:

- Catalogue (dcat:Catalog), defined as *a curated collection of metadata about datasets*
- Catalogue Record (dcat:CatalogRecord), defined as *a record in a data catalog, describing a single dataset*
- Dataset (dcat:Dataset), defined as *a collection of data, published or curated by a single agent, and available for access or download in one or more formats*
- Distribution (dcat:Distribution), defined as representing *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*

The data model of DCAT is presented in Figure 3.

[13] W3C. Data Catalog Vocabulary (DCAT). W3C Recommendation 16 January 2014. http://www.w3.org/TR/vocab-dcat/
[14] W3C. Government Linked Data Working Group. https://www.w3.org/2011/gld/charter
[15] US spelling from the original.

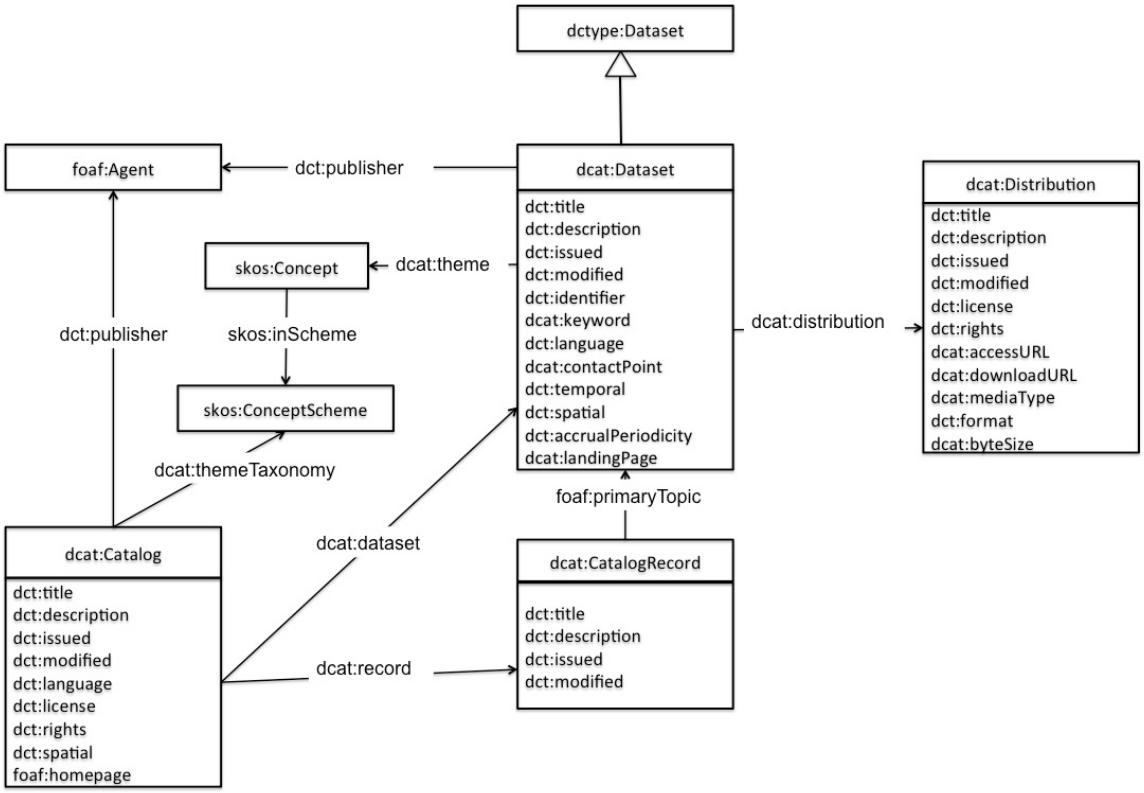

Figure 3: DCAT schematic data model

3.2.2 DCAT-AP for open data portals in Europe

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)[16] of the Application Profile was published in September 2013. In 2015, a revised version (1.1)[17] was developed and published in November 2015 with changes based on requests from implementers of the first version.

The data model of DCAT-AP is presented in Figure 4.

[16] European Commission. Joinup. DCAT application profile for data portals in Europe. Final. https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-application-profile-data-portals-europe-final

[17] European Commission. Joinup. DCAT application profile for data portals in Europe. DCAT-AP v1.1. https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11

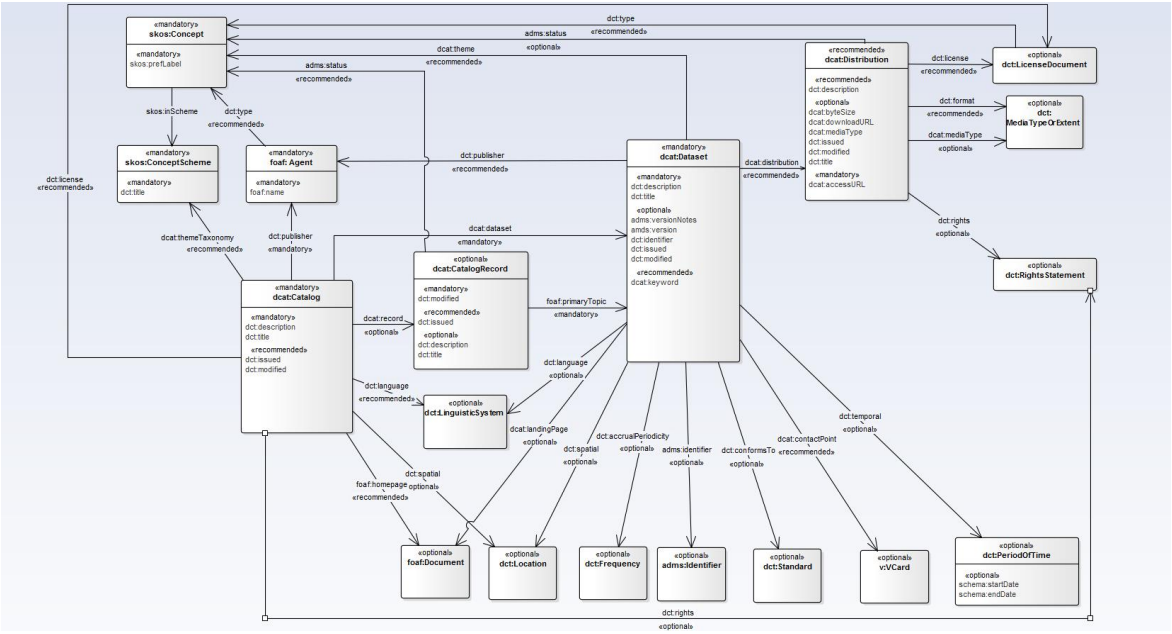

Figure 4: DCAT-AP Data Model

3.2.3 GeoDCAT-AP

GeoDCAT-AP is an extension of DCAT-AP for describing geospatial datasets, dataset series, and services. It provides an RDF syntax binding for the union of metadata elements defined in the core profile of ISO 19115:2003 and those defined in the framework of the INSPIRE Directive. Its basic use case is to make spatial datasets, data series, and services searchable on general data portals, thereby making geospatial information better searchable across borders and sectors. This can be achieved by the exchange of descriptions of data sets among data portals.

GeoDCAT-AP is a joint initiative of the Joint Research Centre (JRC), Unit H.6 (Digital Earth and Reference Data), the Publications Office of the European Union (PO), and the Directorates-General for Informatics (DIGIT, in the context of the ISA Programme) and Communications Networks, Content & Technology (CONNECT) of the European Commission. More than 52 people from 12 EU Member States contributed to the specification in the Working Group or during the public review period.

The first version (1.0)[18] of the GeoDCAT-AP was published in December 2015.

3.2.4 The Data Cube Vocabulary

To be completed

[18] European Commission. Joinup. DCAT application profile for data portals in Europe. GeoDCAT-AP v1.0. https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/geodcat-ap-v10

4 USE CASES

4.1 Improving the quality of the metadata of statistical datasets on open data portals

Within the EU, Eurostat is the organization having as mission to provide the European Union with statistics at European level that enable comparisons between countries and regions.

In February 2015, Eurostat published more than 6500 datasets on the European Union Open Data Portal (EU ODP). That represents approximately 81% of the datasets in the European Union Open Data Portal. In practice many of the other datasets on the EU ODP are more elaborated datasets based on the datasets provided by Eurostat. On other governmental open data portals, the quantitative impact of statistical data is similarly high.

So improving the metadata quality by establishing a dedicated extended profiling of DCAT-AP, StatDCAT-AP, for statistical data has an important impact in the already published dataset records. The improvement increases public and cross-sector access to this category of high value datasets.

4.2 Cross-domain integration of data infrastructures

At inter-institutional level, Eurostat plays an important and active role in constantly improving the exchange of statistical data. In the recent past, the world wide most prominent statistical data organizations, including Eurostat, defined and adopted the SDMX standard for the exchange of statistical data. SDMX ensures the exchange of statistical data happens without loss of information, in particular provenance information. Decision making on the sending and the receiving end of the exchange is hence based on the same information.

Open Data Portals are catalogues of dataset metadata descriptions. Within the European Union, the application profile of the W3C standard DCAT, DCAT-AP harmonizes the dataset metadata descriptions. By correlating the metadata descriptions provided by SDMX and other existing standards for statistical data, both worlds get better connected. StatDCAT-AP aims to facilitate a better integration of the existing statistical data portals with the Open Data Portals, improving the discoverability of statistical datasets.

Today Eurostat and Publications Office have established a first version of such integration. This experience and the experience gathered during work to define StatDCAT-AP can be transferred to similar setups in the EU member states.

Note that it is not the objective of StatDCAT-AP to cover actual data. For that the W3C vocabulary DataCube[19] exists. Work on StatDCAT-AP may, however, include discussions at this level since it may improve insight.

[19] W3C. The RDF Data Cube Vocabulary. W3C Recommendation 16 January 2014 http://www.w3.org/TR/vocab-data-cube/

4.3 More use cases…

More use cases to be contributed through the work of the Working Group.

5 METHODOLOGY

5.1 ISA Core Vocabulary process and methodology

This work is conducted according to a process and methodology[20] that were defined for the ISA programme. The process involves the setting up of the Working Group and the publication of drafts of the specification with external review. The methodology is concerned with the elements that the specification should contain, including use cases and definition of terms and vocabularies.

The objective of the process and methodology is to involve the main stakeholders and to reach consensus in an open collaboration.

The work is conducted in a transparent manner, visible to the public through:

- A Web page
  https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/description
- An issue tracker
  https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/issue/all
- A mailing list
  http://joinup.ec.europa.eu/mailman/listinfo/stat_dcat_application_profile

5.2 Analysis and decision framework

Outlining principles, e.g. aligning with DCAT and DCAT-AP, focusing primarily on metadata elements that contribute to discovery, using metadata terms from existing, well-known and well-maintained vocabularies, including ISA Core Vocabularies, encouraging/mandating use of common controlled vocabularies, preferable ones maintained in MDR by Publications Office, finding appropriate balance between simplicity and complexity from the perspective of the widest, non-specialist audience etc.

5.3 Stakeholders

List of participating organisations. To be added.

5.4 Time plan

December 2015: invitations to stakeholders, set up collaboration infrastructure

January 2016: collect requirements and suggestions

5 February 2016: Familiarisation Webinar

February 2016: first draft based on initial analysis and issues raised

[20] European Commission. Joinup. Process and methodology for developing semantic agreements. https://joinup.ec.europa.eu/community/core_vocabularies/document/process-and-methodology-developing-semantic-agreements

11 March 2016: first virtual WG meeting to discuss first draft

March 2016: second draft based on discussions and decision in the WG

Late March 2016: second virtual WG meeting to discuss second draft

April 2016: third draft, preparing for public review

Late April 2016: third virtual WG meeting to discuss final draft for public review

May and June 2016: public review period

July 2016: fourth virtual WG meeting, discuss and resolve public comments received, publication of StatDCAT-AP

6 OVERVIEW OF STATDCAT-AP DATA MODEL

6.1 UML Class diagram

Diagram.

6.2 Description of classes

Listing of all classes in the model, with an indication what they mean in the context of statistical (meta)data and whether they are mandatory, recommends or optional. The intention is that the classes in StatDCAT-AP respect the rules in the general DCAT-AP, in order for StatDCAT-AP to be conformant to DCAT-AP.

6.3 Namespaces

Table of namespaces used in the data model.

7 EXTRACTION GUIDELINES

Table under development at https://docs.google.com/spreadsheets/d/1JPM0gM-MkM7o4FLItLczXWMR-R1-fRlTvExfkE6oYlA/edit?pref=2&pli=1&hl=en#gid=0.

8 CONTROLLED VOCABULARIES

List of controlled vocabularies to be used in StatDCAT-AP descriptions, as much as possible applying vocabularies used in the general DCAT-AP, augmented by appropriate vocabularies of common usage in the statistical domain.

9 CONFORMANCE STATEMENT

Describing minimum requirements.

10 ACKNOWLEDGEMENTS

Table with all participants with affiliation.

REFERENCES

There are no sources in the current document.