Towards an open government data ecosystem in Europe using common standards

Wendy Carrara, Makx Dekkers, Benjamin Dittwald, Simon Dutkowski, Yury Glikman, Fabian Kirstein, Nikolaos Loutas, Vassilios Peristeras, Brecht Wyns

This study was prepared for the ISA² Programme by:

PwC EU Services, Capgemini Consulting, Fraunhofer Fokus, AMI Consult and Vassilios Peristeras (International Hellenic University, Thessaloniki, Greece)

Publication date: 06 June 2017

Disclaimer:

Contents

List of Figures

List of Tables

1. INTRODUCTION AND CONTEXT

Governments have a large number of basic data which can be of economic and social value to society as a whole. Along those lines, more and more European countries are developing policies to release this data as Open (Government) Data. Open Data refers to information that can be freely used, modified, and shared by anyone for any purpose. It must be available under an open licence and provided in a convenient and modifiable form that is machine readable[i].

The benefits of Open Data[ii] are diverse and range from improved efficiency of public administrations and economic growth in the private sector to increased government transparency and accountability and general wider social welfare. Open Data improves the efficiency of public services. Greater efficiency in processes and delivery of public services can be achieved thanks to cross-sector sharing of data, which can for example provide an overview of unnecessary spending. The economy can benefit from an easier access to information, content and knowledge in turn contributing to the development of innovative services and the creation of new business models. Social welfare can be improved as society benefits from information that is more transparent and accessible. Open Data enhances collaboration, participation and social innovation. The direct market size of Open Data in the 28 EU Member States, Iceland, Liechtenstein and Norway (EU28+) is estimated at 75.7 bn EUR in 2020[iii]. In 2020 almost 100,000 jobs based on Open Data are created in the EU28+[iv].

In 2013, the G8 summit defined the importance of Open Government Data by creating the Open Data Charter. This charter emphasises the role that Open Data can play in both governance and growth stimulation. The charter defines five principles that nations that open up their data should follow[v].

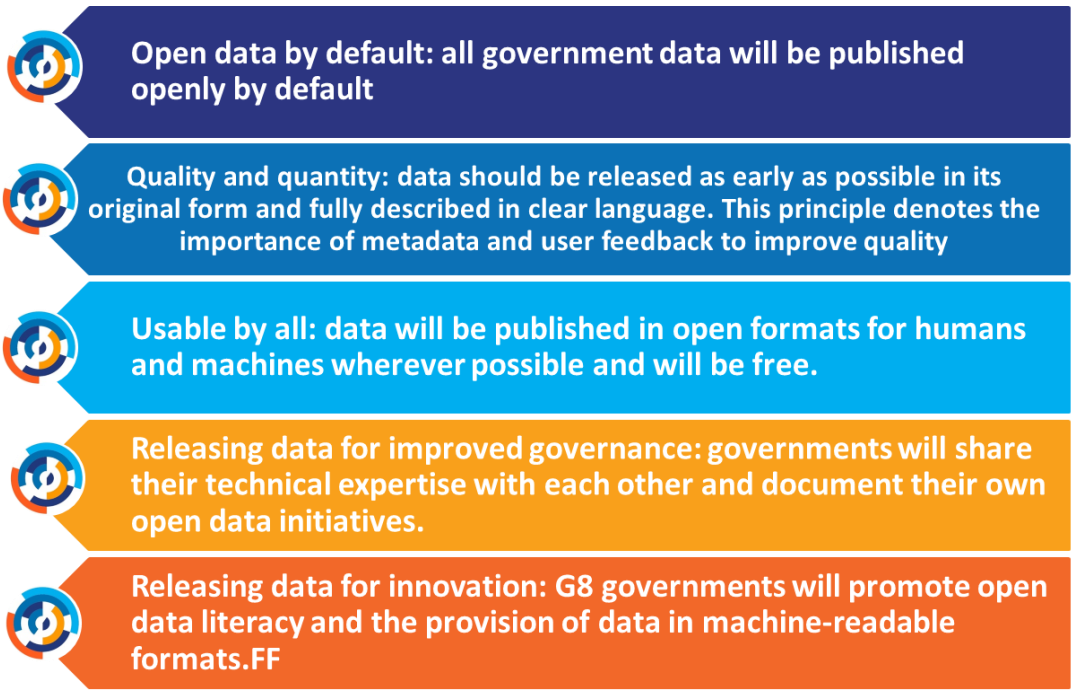

Figure 1: The Open Data Charter Principles

A decade ago, in 2003, the European Union (EU) adopted legislation to foster the re-use of Public Data in Member States via the Public Sector Information (PSI) Directive

2003/98/EC[vi]. The main objective was to ensure equal treatment of all potential re-users where the public sector body had released information for re-use. A revision of the PSI Directive was introduced in 2013 (Directive 2013/37/EU[vii]). The main amendments are the adoption of the "open by default" principle, the breakaway from cost-based charging for PSI towards a marginal cost-oriented fee and increased transparency regarding calculation of the fees, the inclusion of certain cultural institutions as public sector bodies (previously outside the scope), and support to machine-readable and open formats. The European Commission also named five priority domains for release, as not all data sets have been considered as having the same potential for re-use[viii]. Geospatial data, earth observation and environmental data, transport data, statistical data and company data (e.g. business registers) are recognised as having the highest re-use value. All the EU countries with very few exceptions have completed the transposition of the revised PSI Directive.

A recent study on Open Data Maturity[ix] shows that although substantial differences exist between countries, European countries have made clear progress on their Open Data journey. Results indicate that the majority of the EU28+ countries have successfully developed a basic approach to address Open Data (as measured by **Open Data Readiness**[x]). The overall Open Data Maturity groups countries into different clusters: Beginners, Followers, Fast Trackers and Trend Setters (Figure 2).

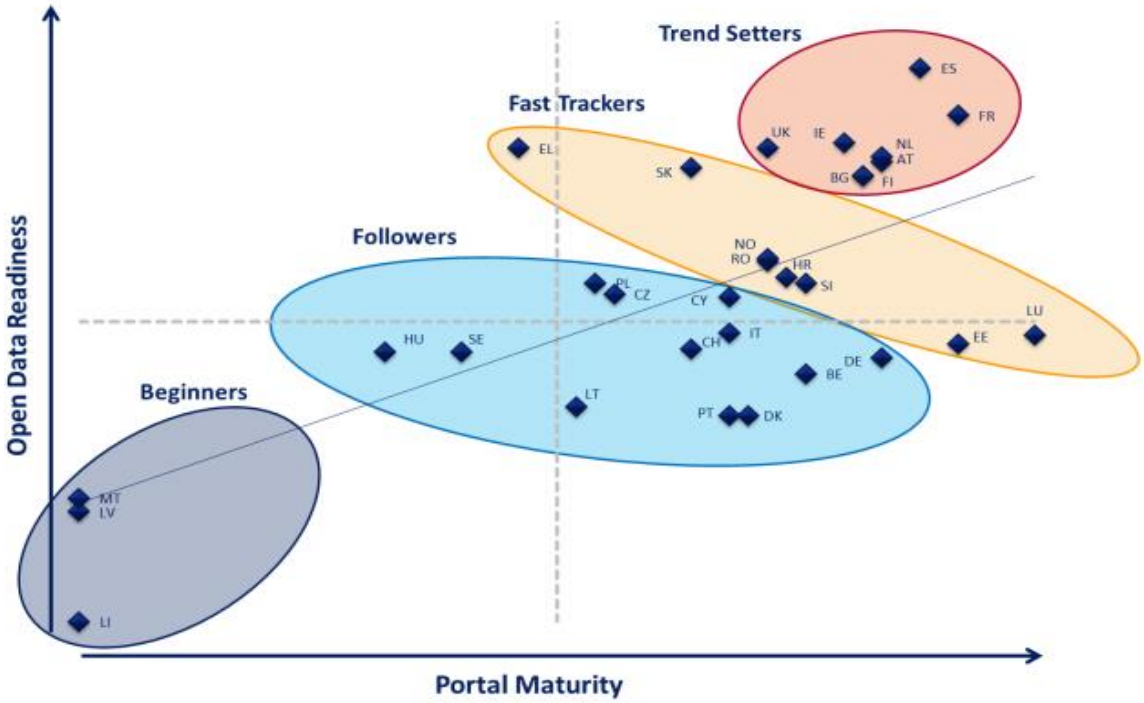

Figure 2: EU28+ Open Data Maturity clusters

2. THE PROBLEM: OPEN DATA PORTALS FRAGMENTATION

In Europe, the different data actors, listed below, form groups with different needs and capabilities, these groups compose an ecosystem where the needs from one actor are answered by the capabilities of others. The connections are within and between groups.

- Open government Data providers are public actors possessing datasets for which they include a description on one or more data portals, so that the datasets can be found more easily;
- Open data portals are online platforms which 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.
- Metadata brokers, such as the European Data Portal, facilitate the collection and exchange of description metadata between data portals by ensuring conformance to a common metadata language. They provide metadata harvesting, transformation, validation, harmonisation, publication services, translation of datasets and other services; and
- Data consumers use data portals of their choice to search through various collections of datasets. The data portals allow the user to explore, find, identify and select the datasets coming from different data providers. Data consumers can also be systems (machines). Many different types of data consumers exist such as academia, media-journalists, NGOs or citizens willing for example to improve transparency or to add value to their services by combining data.

We focus below on the open data portals.

In response to the requirements of the revised PSI directive[xi], European public administrations have set up cross-domain and horizontal open data portals. Open Data portals started as collections of datasets but have gradually become focal points for open data initiatives promoting relevant tools, and applications. Even more importantly, the open data portals create communities by establishing a bridge between data providers and data reusers through collaborative and user-friendly online platforms and tools. The Portal Maturity level[xii] as measured by the European Commission in the context of the European Data Portal increased from 41.7% to 64.3% thanks to the development of more advanced features on country data portals. Open data portals have consequently contributed to the establishment of the necessary foundation for a European open data ecosystem. But limitations appeared soon.

In addition to the inherent political and cultural diversity in Europe and to multilingualism, the development of open data portals has not always been coordinated within or across countries. These initiatives often start with a limited local scope i.e. a city or a region and no attention was given to connections with other portals or to compliance with standards.

Open data portals have been based on different products. Currently, the market leader in Europe is CKAN with some variants, followed by other solutions such as Socrata, DKAN, OpenDataSoft, etc. Even if the basic semantics used by these various

products to describe the published datasets have been coherent, over time the differences have grown. Consequently, this has resulted in a fragmented landscape of open data portals as disconnected information islands, making it hard to exchange metadata between them. This situation leads to duplication of information and inconsistencies. It prevents cross-portal search and discovery of datasets. Gradually and as seen from a European perspective, a babel tower of over 150, non-interoperable open data portals has been created. The main drawback from this situation is the fact that European citizens and businesses cannot access from one single point and search for information and data that exists "somewhere out there" in the plethora of open data portal which are scattered in all European countries. This problematic situation has been raising additional obstacles to open data reaching its full potential.

Some of the key challenges include:

1. **Different metadata specifications, i.e. different semantics**: open data portals describe datasets and data catalogues using different metadata standards and vocabularies, often without defining global identifiers for datasets.
2. **Heterogeneous data formats**: data is published in different formats and following different processes, depending on the objectives pursued and the priorities set by the publishers.
3. Different **quality of the published data**.
4. **Various non-interoperable technologies, software platforms and tools**: the majority of open data portals is based on variants of CKAN, but other options also exist, e.g. OpenDataSoft, DKAN, Microsoft SharePoint. There are also custom-made solutions and we can assume that in the future new platforms will be created also by communities and the market. Situation gets complicated when metadata has to be exchanged between portals which are based on different technologies.
5. **24 different languages**: the majority of these portals provides content only in their national language(s).
6. **Various, or no, licences**: on one side, the types of licence attached to datasets are varying between data portals, and are often non-interoperable; on the other side, the description of the licences itself differs in the metadata: some data portals describe the terms and conditions of use while others provide the name of the licence or the URI of the licence.
7. Lack of **awareness on both the publisher- and the user-side**.

In the European environment, having multiple technologies, platforms and languages is not considered a limitation but rather an inherent and essential feature. The European open government data challenge could be summarised as follows: in compliance with the subsidiarity principle, Europe needs to preserve pluralism in decisions for different technological solutions, platforms, processes and languages while making firm steps towards the creation of an open data European ecosystem. This ecosystem should ensure easy access to open government data from all European countries to all European citizens and businesses.

An important use case implied in the description of the European data environment is twofold. First, data consumers want a single point of access which would allow them to search for data across EU Member States, different portals and different organisations from the data portal of their choice. Second, cross-portal search and discovery of datasets hosted on "local" data catalogues from a European single point of access would only be possible if the different portals with different descriptions of metadata would adhere to a common metadata language. These ideas are discussed in the next section.

Interestingly, similar challenges exist inside national administrations: especially in countries with decentralised systems, it is difficult for central authorities to impose the same technological decisions to local and regional authorities. This results in an open data portal babel even inside the boundaries of one country.

3. LINKING EUROPEAN OPEN DATA PORTALS WITH A COMMON LANGUAGE

In this section, the metadata agreement (DCAT-AP) and the conceptual architecture for the creation of a European Data Portal as a single access point for open data in Europe are presented.

3.1. Defining the common language

The use of common standards as a means to increase reuse of data coming from diverged platforms and systems[xiii] is a way to deal with the European open data diversity while respecting the subsidiarity principle and the freedom of technology choice. This article is focusing on a European effort for standards-based harmonisation of dataset and data catalogue specifications in Europe for increasing the interoperability of data portals to create a European data ecosystem. In this way, several of the aforementioned challenges can be overcome.

At the implementation level, the solution implemented in Europe is a federation of data portals. It consists in a thin layer of common metadata standards applied by the multiple data portals included in the federation. It allows cross-portal searches and as a consequence, it improves the data discoverability and the value for the data consumers[xiv].

The challenge of such a metadata harmonisation initiative consists in consensus building and the convergence of opinions and approaches between the different data portals about the details of the respective semantics they should use. Agreement on a common metadata standard will enhance the potential of the data portals to achieve better coordination and interoperability and ultimately end in increased opportunities to share and re-use metadata[xv], thus also reducing metadata creation and management costs. Moreover, it will decrease the software lock-in risks, as it will be made easier to transfer a collection of metadata from an open data catalogue implemented using one technology to another one implemented using a different technology, and will enable the implementation of cross-portal dataset search scenarios, like the one discussed in section 6.

A number of standardisation and harmonisation efforts have been undertaken to increase the discoverability of data. The DCAT Application profile for data portals in Europe (DCAT-AP) is one of them. It is a specification based on W3C's Data Catalogue vocabulary (DCAT) for describing metadata of public sector datasets in Europe. Other specifications for describing datasets include Schema.org[xvi], VoID[xvii] and the CKAN Metadata Schema[xviii]. Some existing initiatives were developed for specific domains, such as ADMS[xix] for describing interoperability assets, SDMX[xx] for statistical datasets, the INSPIRE[xxi] Metadata Schema for geospatial information and CERIF[xxii] for research data.

In the European context of fragmentation and need to share and reuse information governments possess in quantity, the European Commission asked to prepare and define an Application Profile that can be used for the exchange of descriptions of datasets across domains and among data portals. This Application Profile should answer to the need for common metadata standards at the European level. It was developed under the Interoperability for European Public Administrations (ISA²)

Programme of the European Commission, more specifically under its action on promoting semantic interoperability amongst the European Union Member States (SEMIC)[xxiii].

The DCAT-Application Profile is intended as a common layer for the exchange of metadata for a wide range of dataset types. It provides the basic description for open data and open data catalogues. The availability of such a common layer – as a common denominator – creates the opportunity for professional communities to hook into 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 domain or country.

The Application Profile discussed here is based on the specification of the Data Catalog Vocabulary (DCAT) developed initially at the Digital Enterprise Research Institute in Ireland[xxiv] and became later a W3C recommendation under the responsibility of the Government Linked Data Working Group[xxv]. DCAT is a RDF[xxvi] vocabulary designed to facilitate interoperability between data catalogues published on the Web. Additional classes and properties from other well-known vocabularies are re-used where necessary.

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. In this specification, data portals must accept incoming data and transparently provide these data to applications and services. It does neither imply nor prescribe what applications and services finally do with the data (parse, convert, store, make searchable, display to users, etc.). The Application Profile is intended to facilitate data exchange and therefore the classes and properties defined in the specification are only relevant for the data to be exchanged; there are no requirements for communicating systems to implement specific technical environments. The only requirement is that the systems can export and import data in RDF in conformance with this Application Profile. As a result, the autonomy of individual portals is assured and lock-in is avoided.

3.2. The DCAT-AP

The mandatory classes are the following ones: agent, category, category scheme, catalogue, dataset, literal, and resource. In the DCAT-AP specification, any entity is described by a usage note, a URI, and a reference for further details. Table 1 contains the different information provided for the mandatory classes in the DCAT Application Profile for data portals in Europe Version 1.1[xxvii] .

Table 1: Information specified for the mandatory classes

Class name	Usage note for the Application Profile	URI	Reference
Agent	An entity that is associated with Catalogues and/or Datasets. If the Agent is an organisation, the use of the Organization Ontology[xxviii] is recommended.	foaf:Agent	http://xmlns.com/foaf/spec/#term_Agent , http://www.w3.org/TR/vocab-org/

Category	A subject of a Dataset.	skos:Concept	http://www.w3.org/TR/2013/WD-vocab-dcat-20130312/#class-category-and-category-scheme
Category scheme	A concept collection (e.g. controlled vocabulary) in which the Category is defined.	skos:ConceptScheme	http://www.w3.org/TR/2013/WD-vocab-dcat-20130312/#class-category-and-category-scheme
Catalogue	A catalogue or repository that hosts the Datasets being described.	dcat:Catalog	http://www.w3.org/TR/2013/WD-vocab-dcat-20130312/#class-catalog
Dataset	A conceptual entity that represents the information published.	dcat:Dataset	http://www.w3.org/TR/2013/WD-vocab-dcat-20130312/#class-dataset
Literal	A literal value such as a string or integer; Literals may be typed, e.g. as a date according to xsd:date. Literals that contain human-readable text have an optional language tag as defined by BCP 47[xxix].	rdfs:Literal	http://www.w3.org/TR/rdf-concepts/#section-Literals
Resource	Anything described by RDF.	rdfs:Resource	http://www.w3.org/TR/rdf-schema/#ch_resource

The Application Profile is also defining what the mandatory, recommended and optional properties are per class. For example, for the agent class, a mandatory property is the name of the agent and a recommended property is the type of publisher the agent represents.

3.3. The DCAT-AP governance and revision process

The creation process of the DCAT-Application Profile and the revision process of the following specifications were developed by the ISA² Programme of the European Commission. Four groups with specific roles were identified at different governance level:

1. (ST) Steering Committee (ISA Coordination Group, PSI Expert Group (DG CNECT) where member states were represented);
2. (GC) Governance Committee (ISA² Programme Management Team);
3. (OT) Operational Team (Contractor of ISA² action on semantic interoperability);
4. (WG) Working Group (Organisations implementing the specification and Individual experts, led by a chairman and an editor).

The revision process is composed of five phases[xxx]:

- Request handling. This phase starts with the receipt of requests for change (RFC) from stakeholders. A log of all change requests received will be made available online via Joinup. The requests are evaluated by the Operational Team (OT) and grouped into issues on Joinup. Based on the analysis by the OT, the Governance Committee (GC) decides on the further process. If the request is rejected because it is not clear or not relevant for the specification at hand, the GC informs the submitter of the rejection with a justification. If the request is accepted, the GC will schedule the request for inclusion in a new release.

- Request resolution. In order to resolve the requests for semantic changes, the GC establishes a Working Group (WG). The WG elaborates one or more drafts of the revised specification and discusses these drafts until consensus is reached. It then submits the draft to the GC who publishes the draft for public review. The WG resolves any comments and finalises the new specification. The process continues with the Release preparation phase.
- Release preparation. The GC instructs the OT to prepare the specification and any additional documentation. The GC notifies the Steering Committee (SC) that the new release is ready for publication and requests endorsement by the SC.
- Release endorsement. The SC discusses the new release and endorses its publication.
- Release publication. Following endorsement by the SC, the GC publishes the new release and notifies the stakeholders and the wider public of its availability. The new release of the DCAT-AP will be made available on Joinup.

3.4. The DCAT-AP extensions: Geo- and Stat- DCAT-AP

In 2013, the G8 Open Data Charter highlighted the importance of "high value datasets"[xxxi]. In its implementation, statistical and geospatial information were identified as two of the thematic categories among those "those in highest demand from re-users across the EU"[xxxii].

In parallel, it became clear that the geospatial and statistics communities had already shown a great interest on publishing open data relevant to their domain. However, this usually takes place through dedicated geo- and statistical portals, creating once again a multiplicity of data sources, adding to the fragmentation of available open data for the final user who does not really care whether the relevant data come from geoportals, statistical databases or open data portals of general purpose. For bridging the gap between the different worlds i.e. geographic data, statistical data and "general" open data, two specifications, StatDCAT-AP and GeoDCAT-AP[xxxiii], were developed in a fully conformant way with DCAT-AP version 1.1.

GeoDCAT-AP[xxxiv]

The motivations for specifying the GeoDCAT-AP was to enable a cross-domain data portal search for datasets, as for the DCAT-AP specification. More specifically and as already explained, GeoDCAT-AP would facilitate the sharing of descriptions of spatial datasets between spatial data portals and general data portals, and thus help increase public and cross-sector access to such high value datasets.

For this, the objective of the GeoDCAT-AP was twofold:

1. Provide a DCAT-AP-conformant representation of geospatial metadata; and
2. Provide an as much as possible comprehensive RDF-based representation of geospatial metadata, based on widely used vocabularies (as DCAT-AP), trying, at the same time, to avoid semantic loss and to promote cross-domain re-use.

The full alignment with DCAT-AP and other standards like ISO 19115 and INSPIRE Metadata Regulation represented a guarantee for data portals not to be locked in one semantic metadata standard.

StatDCAT-AP[xxxv]

This specification 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 these data.

This work represents a first set of activities 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.

An important note on the scope and limitations of the DCAT-based harmonization. We focus on how challenges of metadata exchange across fragmented data portals can be overcome by adopting a commonly agreed metadata standard for describing datasets: DCAT-AP. However, we need to clarify what is the scope and at the same time the limitation of this metadata-based harmonization. Fragmentation in the open data landscape is not only observed at the level of the metadata, but also in how the data itself is described. The DCAT-AP does not address this level of the fragmentation problem, as it treats the content of datasets as a black box. Therefore, in order to improve the potential of datasets to be combined and to produce value-added services based on Linked Open Data technologies, there is a need to also harmonise the data models used in different datasets. As an example of work in this area, the ISA² Programme of the European Commission addresses this challenge by developing "simplified, re-usable and extensible data models that capture the fundamental characteristics of an entity in a context-neutral fashion": the Core Vocabularies[xxxvi].

4. IMPLEMENTING THE DCAT-AP AT EUROPEAN AND NATIONAL LEVEL

Since its first release in 2014, DCAT-AP has been implemented in several data portals at local, regional, national and European level. As part of an initiative of the ISA² Programme of the European Union, the real-life implementations of DCAT-AP are continuously identified and tracked[xxxvii]. The work around DCAT-AP is an important contribution of the Programme as 62% of the identified users of ISA² specifications indicated that they are using DCAT-AP. Among those, one third uses the extensions GeoDCAT-AP or StatDCAT-AP. Analysis revealed that most of the projects using DCAT-AP required some type of customisation during the implementation phase. To address this, some implementers have created local extensions for the specification while ensuring compliance to DCAT-AP.

4.1. Implementation at national level: using natively DCAT-AP with extensions in Open Data Portals

The DCAT Application Profile has been implemented by over 15 data portals across Europe, as listed in table 2. Although the initial idea behind DCAT-AP was to create a common metadata language, several implementers use DCAT-AP natively and develop their own data models based on DCAT-AP. These local extensions are further explained below. By building local extensions on top of DCAT-AP, implementers will comply with DCAT-AP, which will allow them to easily integrate their data sets with other national or European data portals.

Table 2: implementers of DCAT-AP

Portal	Level	Location
www.europeandataportal.eu/	Member States, EU	EU
data.europa.eu/euodp/	EU	EU
dati.gov.it	National	Italy
data.gov.be	National	Belgium
opendata.swiss	National	Switzerland
data.gov.ie	National	Ireland
data.overheid.nl	National	The Netherlands
data.gouv.fr	National	France
data.gov.ro	National	Romania
datos.gob.es	National	Spain
data.norge.no	National	Norway
oppnadata.se	National	Sweden
www.opendataportal.at	National	Austria
opendata.vlaanderen.be	Regional	Flanders, Belgium
opendata.brussels.be	Local	City of Brussels, Belgium
data.gent.be	Local	City of Ghent, Belgium
data.kortrijk.be	Local	City of Kortrijk, Belgium
opendata.antwerpen.be	Local	City of Antwerp, Belgium

As already discussed, DCAT-AP provides a core description of open datasets and open data portals. It targets to be cross-border and cross-domain. We have seen in section 3.4 that domain-specific extensions have already been developed for the domains of

statistics and geospatial data. In a similar way, implementations within a national domain may have different and/or additional requirements and therefore may need to define extensions to the basic profile.

In this direction, several EU Member States have therefore extended the specification to meet local requirements:

- **The Netherlands**: DCAT-AP-NL, available via https://data.overheid.nl/IPM-Datamodel
- **Norway**: DCAT-AP-NO, available via https://doc.difi.no/dcat-ap-no/
- **Italy**: DCAT-AP_IT, available via http://www.dati.gov.it/consultazione/dcat-ap_it
- **Switzerland**: DCAT-AP for Switzerland, available via http://handbook.opendata.swiss/en/library/ch-dcat-ap
- **Germany** has announced[xxxviii] in December 2016 that it will develop a new specification for describing the metadata of public sector datasets, DCAT-AP.DE.

During discussions with implementers of DCAT-AP, the following reasons for creating local extensions were identified:

- Translation of labels: the labels included in DCAT-AP are described in English. However, national implementations might require these labels to be encoded in another language;
- Changing cardinalities: in order to allow DCAT-AP to be flexible enough to be implemented in different data portals, many properties are defined as optional. National implementers, however, might prefer to make some of these optional properties mandatory; and
- Ensure compliance with National guidelines and specifications, e.g. the Dutch Information Publication Model (IPM) and its reference data, such as code lists and organization identifiers.

In order to ensure that extensions are compliant with DCAT-AP, some guiding principles[xxxix] have to be taken into account. Any extension needs to respect the minimum conformance requirements as defined in the specification. More specifically:

- Extensions must not widen but may only narrow down the usage notes as specified in the specification, so that all information provided according to the extension remains valid for DCAT-AP v1.1.
- Extensions may add classes that are not specified in DCAT-AP; however, an extension should not add classes that are similar to existing classes.
- Extensions may add properties that are not specified in DCAT-AP; however, an extension should not add properties that are similar to DCAT-AP properties.
- Extensions may change the cardinalities for properties respecting the following rules:
    - Mandatory properties must be mandatory in the extension.

- Recommended properties in may be declared optional or mandatory in the extension.
- Optional properties may be declared recommended or mandatory in the extension.
- Recommended and optional properties may be removed from the extension.

- Extensions must include all the mandatory controlled vocabularies as listed in the specification.
- Extensions may add mandatory controlled vocabularies.

4.2. Implementing the DCAT-AP in the European Data Portal

In this part, we present first the high-level requirement and conceptual architecture for implementing DCAT-AP at the European level and more specifically as the core specification for the European Open Data Portal. Then we provide some more technical details explaining how the federation currently works and which technologies are used. Although this second part is technical, the description remains at a high level and we don't explain implantation details.

4.2.1. *High-level requirements and conceptual architecture*

Implementing the federation of portals implies to develop an architecture that allows separated data actors to pursue their common objectives such as: allowing cross-data portals searches, offering to data portals one appearance to their visitors, identifying duplications and gaps between data and metadata or helping identify best practices in the services proposed. The conceptual architecture[xl] of this federation should be characterised by the following points:

- **A semantic alignment** as already discussed: in Europe, a shared language implemented by many data actors of the European ecosystem; and
- **A conceptual architecture** with clear types of possible relations between the actors of the network.

Furthermore, no hierarchy is required between the actors of the network which means that two individual data portals/metadata brokers can decide to connect or not to each other. This justifies why the federation should only build a thin layer in order to ensure the autonomy and flexibility of the data actors.

It is also important to notice that any data actor of the ecosystem can cumulate multiple roles. A data organisation active at the national level could for example combine data portal and metadata broker services: publishing and storing datasets from national data providers and harvesting metadata from other data catalogues to increase their 'searchability' and discoverability.

In this network, the European Data Portal has a particular status since it acts as a central node at the European level for metadata activities of public administrations[xli].

A federation of data portals answers adequately the needs for better interoperability in cross-portal searches:

- Datasets are stored 'locally' in the data portal which is directly in contact with the data provider. The other portals and metadata brokers only use the catalogue of metadata to have the references, descriptions and locations of the datasets;
- A data consumer can search in his or her own language in one centralised portal which is harvesting individual portals with different languages;
- A single point of access is proposed for identifying and discovering data thanks to a common metadata vocabulary, to common search criteria, etc.;
- Different analysis are possible thanks to a federation of portals. For instance clustering, identifying relationships between datasets and developing cross-country and cross-domain data analysis scenarios;
- Subscription services for data consumers who are interested to be notified of new data being published in certain domains or countries; and
- Catalogue entries are harmonised, simplifying the compliance process of new metadata, enabling automated validation of metadata to take place.

On the downside, a federation containing hundreds of thousands of datasets can suffer from duplicates and inconsistencies, making it difficult for data actors to find what they need. Therefore, a federation of portals ideally has to go beyond the agreement on common metadata in order to collect the potential benefits.

The implementation of a federation of data portals is facilitated by an environment that enables information sharing. Such environment corresponds to the ecosystem Europe is evolving to, characterised by interacting and connected actors with a certain autonomy. The organisation(s) implementing the federation should also support actively the actors of the network to establish data and metadata governance practices by sharing guidelines, best practices and services (e.g. registry capabilities).

In the long run, to expand and integrate an increasing amount of data providers into the European data ecosystem, independent and autonomous data portals and metadata brokers must have a value proposition that convinces the data providers to voluntarily comply with the proposed standards.

4.2.2. EDP high-level technical architecture and underlying technologies

Building a data portal to provide access to all existing open data portals in all Member States across Europe is a huge, complicated and challenging task.

The European Data Portal project started 2015 as part of the Connecting Europe Facility[xlii] (CEF) infrastructure and is scheduled till 2018. Whereas the Beta version of the portal was made available in November 2015, the first version of the portal was released at the beginning of 2016 as a beta and rolled out as version 1.0 in the same year. After another year of development, version 2.0 was released on the 1st of March in 2017. From the very beginning, it was specified to fully support DCAT-AP for the stored metadata. This section describes how we achieved this and what future plans we have.

The European Data Portal makes data available and re-usable across Europe. This is done by harvesting metadata from the Member States' data portals and making them

available and searchable through the European Data Portal. Unfortunately, the format and structure of metadata differ from portal to portal. This problem can be addressed by making use of a single, harmonised model, the DCAT-AP specification, which has been developed by the European Commission.

However, taking into account the overall number of open data portals in Europe, there are still few portals currently using natively DCAT-AP for their metadata descriptions. Moreover, these portals use different languages across Europe. This requires translating the metadata in all languages. This feature is key to enable citizens from any EU member state to effectively use the portal and discover and possibly re-use the data they are looking for. The European Data Portal addresses these challenges with a modular architectural approach.

High-level technical architecture

Figure 3 shows the high-level technical architecture of the European Data Portal.

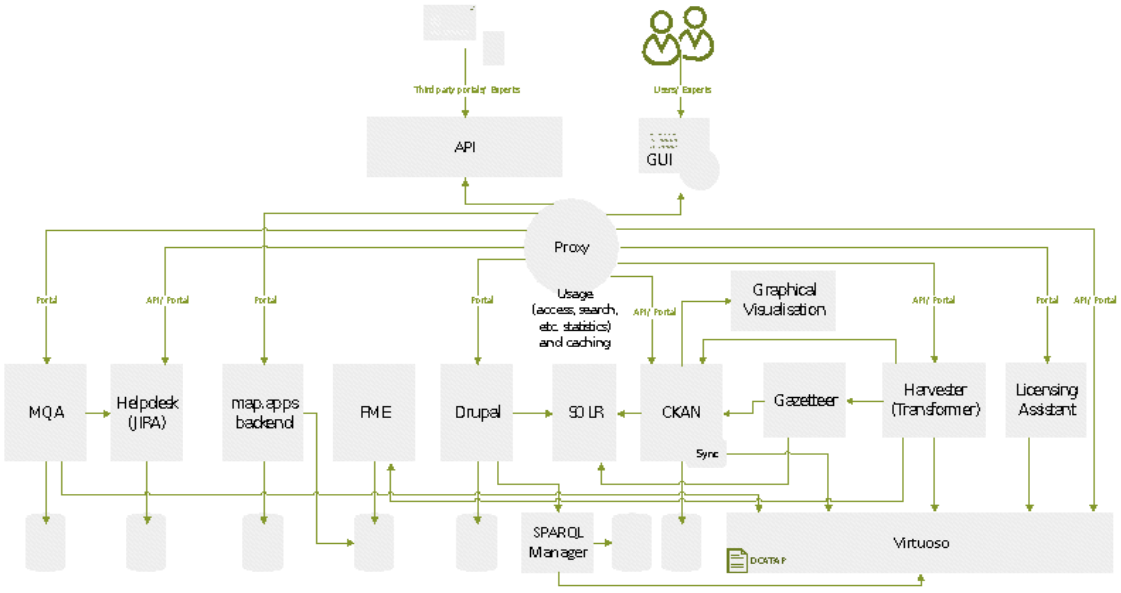

Figure 3: High-Level technical Architecture

The access to the Portal is provided in two ways: a machine-readable API and a human readable web site (GUI). The API enables its users to search, create, modify and delete metadata on the portal. The GUI is basically built on two components: CKAN and DRUPAL. CKAN manages and provides metadata content (including references to datasets) in a central repository. DRUPAL provides the Portal's Home Page with editorial content (e.g. Portal's objectives, articles, news, events, reports, etc.) and links to an Adapt Framework[xliii] based training platform. In addition it offers extended functionalities to registered users via user login. Both systems are used in a side-by-side architecture. A proxy is responsible for delivering the web pages requested by the user. Both systems are equally themed with the same Look&Feel so that the user is not aware on which system he/she is currently browsing.

The Portal GUI supports all 24 official EU languages for main editorial and main metadata content (using the CEF Automated Translation building block powered by MT@EC[xliv]). Training content, requiring human translation, is available in English and

French only. Additional material can be made available in English or in the source language. In terms of search functionalities, the portal uses the SOLR search engine in order to separately search for editorial content in DRUPAL and for datasets in the CKAN repository. The GUI includes a Licensing Assistant component that supports the user by providing legal information on the permitted usage of a specific dataset in terms of licenses that apply to the dataset. The SPARQL Manager component allows the user to enter and run SPARQL queries on the Virtuoso linked data repository. The Virtuoso quad store is used for storing all metadata in DCAT-AP linked data. It also allows the logged-in user to store and re-run SPARQL queries and notifies the user when a query execution provides different results compared to past executions.

Using the map.apps backend application, geospatial data is visualised in the form of geographical maps. Therefore, Web Map Services (WMS) that are made available in compliance with the INSPIRE directive are used. The application is a proprietary solution that comes with different tooling and thematic focus, a graphical configuration interface, supports responsive web-design and internationalisation files. The application also implements the OSGI specification on the client side (in JavaScript) allowing www.europeandataportal.eu sharing and re-usage of the bundled application logic as well as a straightforward maintenance. Statistical data that is linked to datasets can be visualised in tabular (tables) and graphical (charts) form.

On the Harvesting side, the portal follows a two-fold architecture too. CKAN is used as the central metadata repository for storing, browsing and searching datasets in a PostegreSQL relational database. In order to also support a linked data functionality, the CKAN metadata is replicated into a Virtuoso quad store repository via a CKAN synchronisation extension, in order to ensure that both repositories have the same set of metadata. The Harvester is a separate component that is able to harvest data from multiple data sources with different formats and APIs. The harvester is acting as a single point of entry for all metadata that is harvested, transformed into the CKAN JSON schema and pushed into the CKAN repository. The Gazetteer component is used by the Harvester to enhance the metadata with geospatial data and information (geo-coordinates, names, places, etc.). The Gazetteer is mainly used to improve the search functionality for geospatial data. It uses the FME component as a universal spatial ETL tool (Extract-Transform-Load) that supports the accessing, processing and outputting of all spatial file/database formats and that is used for harvesting the sources for geographical names.

The Portal architecture includes three additional components to enhance the quality of the metadata and the portal. A Helpdesk handles user support requests and feedback. The Metadata Quality Assistant (MQA) periodically generates reports on the quality of the harvested metadata. The third component is the monitoring component based on PIWIK and located at the Proxy in the architecture. In the full respect of data privacy, it records requests and user interactions on the portal in order to generate anonymised user traffic statistics that will help enhancing the usage of the Portal.

Currently (March 2017) EDP is harvesting 76 sources, including INSPIRE based (35), the others are CKAN portals (28), file dumps in different formats (4), or portals with more or less proprietary APIs (3). Many datasets are not compliant with our quality requirements, but all sources together summarize to around 640,000 datasets. Most source portals support incremental harvesting; therefore, they can be updated on a

daily base. The bigger sources which are not supporting incremental harvesting are scheduled on a weekly base. There are a few source portals that have a very dynamic repository, which results in several thousand datasets to be updated created or deleted per day. Therefore, the actual amount of datasets in the EDP can vary from day to day.

DCAT-AP Mapping and Implementation

The European Data Portal stores the metadata of each dataset in two separate persistence systems. The first database makes the metadata accessible via the CKAN-based portal and the second database provides the metadata as DCAT-AP compliant linked data in the Virtuoso triplestore. The linked data version represents the complete presentation of each dataset. The underlying technology and data scheme of CKAN is fixed and is based on the relational database PostgreSQL and the search server Solr. CKAN employs a flat key-value based data structure with a predefined set of default fields. This schema can be extended with arbitrary fields for storing custom data. However, it is only possible to have rigid and closed database schemas. In comparison, the linked data methodology of DCAT-AP provides the possibility to describe data fields using an expressive existing vocabulary. Therefore, the data representations of CKAN and the triple store are fundamentally different. This can be seen as a semantic gap which needs to be addressed when mapping DCAT-AP data to CKAN schema. The objective is to reduce this gap in order not to lose any information. This is achieved in three steps:

The DCAT-AP classes have been mapped to appropriate CKAN concepts. E.g. datasets have been mapped to packages and catalogues to organisations. Then, for each class, each property was mapped to a semantic equivalent core field in CKAN. E.g. dct:description to notes. For all properties, which are not covered by CKAN core fields, so-called extra fields were created, for example dct:contactPoint. In general, DCAT-AP covers much more metadata than CKAN. Therefore, more than 25 extra fields were added. In addition, many core fields are equivalent in DCAT-AP. Finally, a complete mapping[xlv] from DCAT-AP to CKAN was created.

For each mapped property, a detailed data structure had to be created. Where Linked Data (RDF) offers a complex, flexible and open data structure, the possibilities in CKAN are limited. It uses basic JSON and hence the JSON data structures, which are limited to numbers, dictionaries, strings and lists. One solution might have been to utilise JSON-LD as a linked data representation. Since this would have led to a fundamental change in CKAN's core technology stack it proved not to be practical. Therefore, a custom mapping was implemented. The main obstacle was that the ranges of most DCAT-AP properties are open. For example dct:contactPoint is defined as vcard:Kind, which defines many valid properties. The solution was to map only the most common properties to the CKAN data structure. Figure 4 illustrates and exemplary mapping.

Figure 4: DCAT-AP to CKAN Mapping

Cardinalities in DCAT-AP are mapped to JSON data structures, e.g. to lists. In addition, linked external resources are stored using URIs.

The schema and data structures created in step 1 and 2 are implemented in CKAN as an extension, which modifies and changes the CKAN core schema. Data that comes in formats other than DCAT-AP can be translated into 100% DCAT-AP compliant metadata. However, due to the openness of DCAT-AP, it may happen that small pieces of information are lost if DCAT-AP is harvested directly.

The synchronisation with the Virtuoso triple store is done on every write operation, no matter whether the updates come from the harvester or via the frontend. The EDP extension for CKAN hooks into the action layer for write, update and delete methods for both datasets or resources and make sure that the same operation is applied to the content of the triple store. Although Virtuoso is used in EDP, the extension does not actually depend on any specific product, since the communication with the triple store is via pure SPARQL, a standardized RDF query language.

The same SPARQL endpoint that is used for the synchronization is also exposed to the public, but without write access. The EDP CKAN, together with the Virtuoso, is configured in such a way that, by appending ".rdf" or ".n3" to the end of the URL of a detail page, CKAN returns the metadata as RDF in the requested format. Additionally, each dataset in EDP also has a unique RDF resource URI, which is resolvable. If a user sends a HTTP get request to this URI with a header asking for HTML, s/he will be redirected to the appropriate CKAN detail page. But if a user asks for RDF (by using the header "Accept: application/rdf+xml"), s/he will get the requested RDF format in the response.

Current challenges and future work

Several challenges have been identified during the implementation of the European Data Portal. Those related to the DCAT-AP have been communicated in the relevant working group and are discussed in the next part.

As a further step in the EDP's development it is planned to support DCAT-AP natively, including Stat- and Geo DCAT-AP. This will ensure 100% DCAT-AP compliance across all harvested metadata and enriches the search functionality.

5. OVERCOMING DCAT-AP IMPLEMENTATION CHALLENGES AND GUIDELINES

Real-life implementations of DCAT-AP, such as the adoption in the European Data Portal and other European data portals as described in the previous chapters, have uncovered a number of challenges when implementing DCAT-AP. This chapter presents the identified challenges and the guidelines that were developed to help implementers overcome those challenges.

In December 2015, the members of the working group that were involved in the revision of DCAT-AP were invited to participate in the identification of implementation challenges and in the development of guidelines. The issues identified during the revision process of DCAT-AP, which led to the publication of version 1.1, were used as a starting point for the working group. In a first stage, the working group members were given the opportunity to identify additional issues. As a result of this exercise, the working group identified four main categories of issues that implementers encountered when implementing DCAT-AP: deployment, mapping, modelling and usage issues.

To enable consideration of those issues that were most interesting to the community, the working group members were invited to vote for the issues and indicate which issues were most interesting from their personal perspective. From the ranking based on the voting, ten main issues were selected for further processing. The most important issues and proposed resolutions are presented in the following sections. As the work on implementation guidelines is ongoing, we refer to Joinup for an exhaustive overview of identified issues[xlvi] and guidelines[xlvii].

The remainder of the issues were recorded and served as input for future work. Some issues require an update of the DCAT Application Profile and will be taken into account in future versions. Other issues, which did not make it through the prioritisation for the first guidelines, are used in the process for developing new guidelines which has started in October 2016 and is currently ongoing. A third category of issues consist of issues that cannot be solved at the level of the Application Profile, but would require an update of DCAT itself, which is managed by the World Wide Web Consortium (W3C).

5.1. Deployment challenges

Deployment issues concern operational approaches, including tools for mapping, export and harvesting. The main challenges faced by organisations when conforming their data portal to the DCAT-AP were related to the compatibility of various tools, the validation of the inputs, the data versioning, and the detection and management of duplicates.

5.1.1. Tools for DCAT-AP

For the first challenge, the role of common or compatible tools for the creation and the maintenance of metadata and for mapping and exporting metadata from local systems to DCAT-AP-compliant metadata was not considered in the development of the DCAT-AP. In order to overcome those issues, implementers have developed their own tools, which are often available as open source software. In order to support the implementation of the DCAT-AP, an overview of the existing tools was built.

Concretely, the various developers who implemented a DCAT-AP solution had the opportunity to add their tool to a list by completing a form available on Joinup[xlviii] . The list of tools contains editors, validators, harvesters and exporters of metadata compliant with DCAT-AP.

Validators

- Open Data Support DCAT-AP Validator[xlix]
- The DCAT-AP validator enables you to check metadata descriptions of datasets for integrity, consistency and conformance against the DCAT-AP specification.
- DCAT-AP Validator for öppnadata.se[l]

Editors

- GeoNetwork opensource[li]
- GeoNetwork is a catalogue application mostly focussing on registration of spatial resources, such as datasets, maps, services, models and software. The goal is to improve discoverability and usability of those resources. The application has options for editing registrations, validation of registrations and harvesting.
  GeoNetwork focusses on the use of standards such as Catalogue Service for the Web, OpenSearch, oai-pmh and supports storage and output schema's such as iso19115, DCAT-AP, schema.org etc.
- EntryScape Catalog[lii]: EntryScape Calog is a collaborative editor for dataset descriptions in RDF according to DCAT-AP. There is also support for import, basic validation, export and an associated data portal that can be used for previewing datasets. National or topical adaptions of DCAT-AP can easily be supported by adapting the metadata templates (RDForms templates). The platform is available as a cloud offering and can, when needed, be installed on-premise.
- Esri Geoportal Server[liii], a catalogue application that supports INSPIRE profiles offering DCAT-AP conversion
- LinkedPipes ETL[liv], a lightweight ETL tool for Linked Data. The tool provides DCAT-AP support in a form of components ready to be used in the tool. They provide a dialog for DCAT-AP metadata for datasets and distributions.

Harvesters

- Geocat CKAN Harvester[lv], a harvester for the GeoNetwork-based Geocat
- DCAT Harvester for CKAN[lvi]: Consume and provide DCAT data via CKAN
- Esri Geoportal Server[lvii], a harvester for DCAT feeds.

Exporters

- CKAN-DCAT[lviii]: a CKAN extension provides plugins that allow CKAN to expose and consume metadata from other catalogs using RDF documents serialized using DCAT. This tool could be further improved to be compliant with DCAT-AP.

5.1.2. Detecting and handling duplicates

The existence of duplicate datasets within and across data portals leads to multiple interoperability-related issues. Since representations of one dataset exist on several portals (due to the federated architecture), it is difficult for a data consumer to identify which is the original source, which might be necessary to identify original licence statements, provenance information, linked data sets, etc.

Many DCAT-AP implementers suffer to deal with the identification and handling of duplicate datasets. Duplicates are specifically a problem when a central data portal or aggregator, for example at a national level, collects datasets from other data portals, for example regional data portals. When the same dataset exists on several regional portals and they are not identified using a unique and stable identifier, it is difficult for the national data portal to automatically identify the duplicate datasets.

Two different types of duplicates can occur when an aggregator harvests descriptions of datasets from various sources:

- In the harvested data, there are two or more descriptions of the same physical data file or API/end point – in this case, the download or access URLs in the descriptions are the same; and
- One or more of the harvested sources describe a copy of the data file or API/end point – in this case, the descriptions refer to different physical files.

The group agreed on a recommendation which consists of three pointers:

- Assign a stable identifier to the dataset in the catalogue where the dataset is first published. This should be the primary identifier of the dataset;
- In the case of duplicates, other locally minted identifiers or external identifiers such as Datacite, DOI, ELI etc. will be assigned to the dataset. As long as they are globally unique and stable, these identifiers should be included as values to the descriptive DCAT-AP property adms:identifier;
- Harvesting systems should not delete or change the value of adms:identifier and only use it to compare harvested metadata to detect duplicates.

5.2. Mapping challenges

Mapping issues are issues that have to do with how local classifications can be mapped to DCAT-AP themes and how DCAT-AP imports can be mapped to existing systems.

5.2.1. Mapping national themes to the "Data Theme" Named Authority List

Members of the working group indicated that a main mapping issue is related to the use of the controlled vocabulary for dataset themes, the Metadata Registry[lix] "data theme" vocabulary[lx]. National implementations may use national classifications for published datasets, partly because such national classifications pre-existed the definition of the MDR Themes Vocabulary, and partly because national services require slightly different themes.

In order to address the issue of different theme categorisations being used at different administrative and geographical levels, the use of the MDR Data Themes on

all levels (local, regional, national, European) is encouraged as it creates coherence. If local or national schemes must be used, mappings to the MDR Data Themes should be made available publicly. The European Data Portal, which is creating mappings between the MDR Data Themes and local or national schemes, and the Publications Office of the EU, being the owner of the MDR Data Themes, should work together in co-ordinating mappings to the MDR Data Themes. The mappings that have already been created should be published on the MDR together with the Data Themes NAL. The latest version of all mappings should always be accessible via the MDR.

The use of a common set of values for data themes, or alignment of different schemes via the creation and publication of mappings to the MDR Data Themes, increases interoperability as it helps datasets published on different catalogues to be classified following a unique and unified classification scheme. This is particularly relevant for cases such as the European Data Portal which aggregates metadata from different catalogues. Moreover, the use of common data themes improves the findability of the categorised datasets via different points of access. A detailed description of the guideline is available on Joinup[lxi].

5.3. Modelling challenges

Modelling issues are related to the semantics of the entities defined in DCAT-AP and their relationships.

5.3.1. Dataset series

During the revision process of the DCAT-AP in 2015, it was noted that the DCAT specification only considers relationships between a catalogue and the datasets described in the catalogue, and between a dataset and the distributions that represent the manifestations of the dataset.

The specification of DCAT was silent on any relationships between catalogues, between datasets and between distributions. However, in real-world implementations, such relationships do exist and may be modelled in different ways. An example of a common relationships of this type is time-series. In some implementations, time-series are modelled as distributions of a single dataset; in others, as separate datasets with or without links between them. The lack of convergence towards a common approach to modelling such relationships in DCAT-AP impeded interoperability among catalogues[lxii] .

DCAT-AP allows relating datasets as 'versions' using dct:hasVersion/dct:isVersionOf but it is not clearly described in which cases to use these properties.

Based on consideration of existing practices and further discussion, the following approaches are suggested:

- If users are mostly interested in the individual members of the series, it is recommended to describe them as separate datasets. While DCAT itself and the DCAT-AP do not specify a mechanism to express the relationship among such datasets, the GeoDCAT Application Profile proposes one.
- If users are mostly interested in the series as such, it is recommended to describe the members as multiple distributions of a single dataset. In order to provide information about the coverage of the distributions, the metadata for

the distributions may include temporal or spatial coverage (dct:temporal and dct:spatial) to assist users to navigate to a particular file within the collection.

- If user expectations are difficult to determine, creating separate datasets and one combined dataset with the members as distributions is recommended.
- If you want to indicate precedence/sequence among different versions of a data set, the DCAT-AP proposes the use of dct:hasVersion/dct:isVersionOf. Moreover, a versioning scheme should be put in place and version numbers should be assigned as value to owl:versionInfo. adms:versionNotes can be used for describing the differences between the current version and the previous one, or for indicating that a newer version is more valid than an older one.

In the absence of consensus on how to model temporal or spatial series, the recommendation intends to give advice that considers the issue from the user perspective and may lead to a more coherent environment that is understandable to users, while retaining flexibility in the approach followed by data providers.

5.3.2. *Provenance*

Another modelling issue faced was related to the provenance of the metadata[lxiii] . The DCAT model treated the descriptions of datasets in a catalogue as entities that only exist in the context of the catalogue, and did not consider situations where these descriptions are imported from and exported to other catalogues.

In an environment where descriptions of datasets are exchanged among data portals, the situation that DCAT-AP is designed for, it may be important for users to understand where data comes from and how it may have been modified along the way. For example, it could support credibility of a dataset to know which organisation created the metadata for it and how the description was modified along a chain of exchanges.

DCAT-AP specifies an optional property dct:provenance for dataset but does not provide any guidance on how to describe instances of the class dct:ProvenanceStatement.

As the provision of provenance information is not wide-spread between the national implementations and information in free text does not allow further processing, the usefulness of such information in (international) harvesting is questionable and the information may be ignored. Local implementations are of course free to provide provenance information satisfying local requirements. In the absence of commonly agreed approaches, ignoring provenance information does neither help nor hinder interoperability.

5.4. Usage challenges

Usage issues are issues that require further clarification as to the use of DCAT-AP properties and classes in practical environments. The biggest challenges regarding usage of the DCAT-AP concerned guidelines and common approaches to increase interoperability.

5.4.1. Licence documents and licence URIs

Specific advice would help implementers to choose the right way of expressing licences and increase interoperability across implementations. DCAT-AP implementers were applying different practices for describing licences[lxiv] :

- Licences are often referred to the use of the licence name as a free text field or as a URI;
- Both well-known international licences, such as Creative Commons or Open Data Commons, as well as custom national licences are often used;
- In very few cases, terms and conditions are described as free text; and
- Licence types are not commonly provided.

Licences should always be identified with URIs, which should resolve to the description of the licence. Well-known licences should be used wherever possible. If a local or national licence is used, its description should link to a well-known licence on which it is based.

Moreover, in order to foster the sharing and reuse of government data, it is important for a data provider to clearly specify at which terms and conditions his datasets can be reused. This can be easily done by referring to well-known licences and identifying them using URIs.

5.4.2. Identifiers for datasets and distributions[lxv]

RDF-based implementations of DCAT-AP do necessarily assign the identifiers for the graphs that contain the dataset and distribution descriptions. In these cases, usually the graph identifier of the dataset description is copied into dct:identifier.

Implementations that are not based on RDF need to export descriptions from a non-RDF system to RDF. In some cases, this is done by assembling a single RDF/XML structure that embeds all metadata for the catalogue. Such approaches may embed the descriptions of distributions within the descripting of the associated dataset and embed the descriptions of all dataset in the description of the catalogue, creating a large file that holds all metadata. Such an approach does not require assignment of URIs to the entities, and such implementations may indeed not do that.

In addition, in RDF-based implementations some entities may be modelled as blank nodes, for example a Period of Time may be expressed as a blank node with properties for start and end date. Some tools have difficulties processing such blank nodes.

As a consequence, the following approach was recommended:

- Stable URIs should be minted for all entities;
- If possible, URIs should resolve to metadata (303 redirect);
- URIs generated on export must be unique and stable (same URI every time it is generated);
- Depending on the format of the data (RDF/XML, JSON-LD) specific URIs are defined for each of the entities;
- If necessary, blank nodes to be assigned Skolem URIs[lxvi] ;
- Dataset URI should be copied into dct:identifier.

If stable identifiers are assigned to all entities, the processing of the information will be made easier.

5.5. Identified issues that require a DCAT revision

Although the guidelines described above help stakeholders to overcome implementation challenges, some challenges would need a revision of DCAT.

5.5.1. 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:

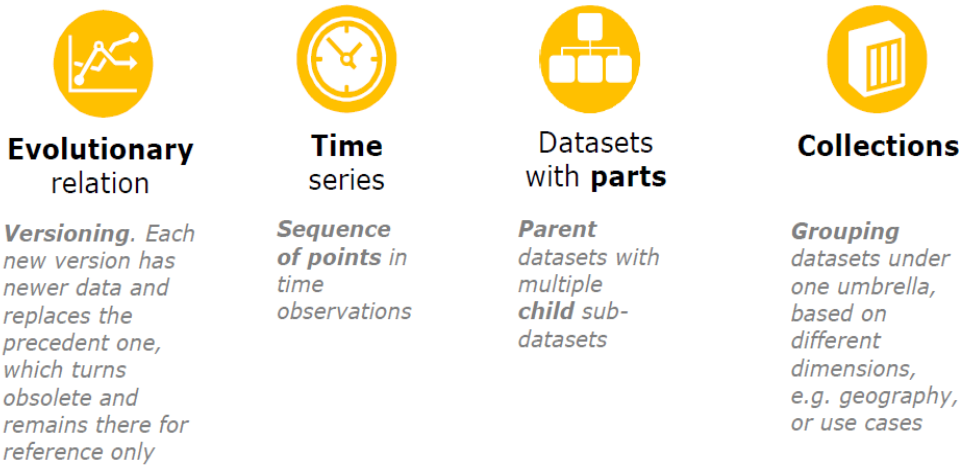

Figure 4: types of dataset relationships

One of the DCAT-AP guidelines developed in 2015[lxvii] 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.

5.5.2. 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 it 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 that favour the latter interpretation[lxviii].

In the current situation, a variety of approaches can be observed. In an analysis of the data in the DataHub[lxix] at least five different approaches could be observed.

Although it may be too late to try and create a consistent approach given the existing landscape, 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 the datasets.

5.5.3. Non-file distributions

It turns out that many datasets in the wild are not published as files but can be accessed through 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. For example, specific information is necessary to access 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, in as far as machine-processability is concerned.

5.5.4. 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 would require the package format to be expressed in dct:format of dcat:mediaType but it might also be helpful for an application to know what type of files 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[lxx].

5.5.5. 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.

6. EVALUATION AND BENEFITS

In order to evaluate the success its specifications, the ISA² Programme of the EU has developed a cost and benefits framework for interoperability solutions[lxxi]. The sections below describe the main benefits of DCAT-AP according to this framework.

A benefit of an interoperability solutions is defined as a "concrete distinct advantage or profit that is measurable and can be demonstrated to derive, directly or indirectly, exclusively from the interoperability aspect of a given solution"[lxxii].

6.1. Types of benefits

The work on the DCAT-AP and the activities of the European Data Portal lead to increased awareness among data providers and users on the utility of common standards for metadata, as discussed in chapter 3. Raising the awareness happened in two stages. First, continuous promotional activities for DCAT-AP have enlarged the federation with data portals using the common standard. In a second stage, an increasing number of data providers and users are voluntarily enriching the European open government data ecosystem with standards and supporting tools. The activities conducted by the European Data Portal and the ISA² Programme have proven to be effective, given

- The high number of implementers of DCAT-AP as listed in Chapter 4;
- The large ecosystem of Data Portals feeding data into the European Data Portal;
- The large amount of contributions from DCAT-AP users to the ecosystem, including tools, guidelines and advice for the future development of the application profile.

Reality shows that a willingness to collaborate exists between the data actors in Europe. Moreover, it shows that increasing the discussion opportunities about standards helps organisations consider interoperability as an important factor for open data utility. The large adoption of the DCAT-AP brings multiple benefits to the actors of the open data ecosystem, as summarised in table 3 and further explained in the sections below.

Table 3: Main characteristics of the benefits for DCAT-AP and EDP activities

Type	Stakeholder concerned	Monetisable?	Direct or indirect measures	Time evaluation
Financial				
Reduce operational costs	Data providers Data portals	Yes	Direct	Ex-ante and Ex-post
Vendor lock-in avoidance	Data providers Data portals	Yes	Direct	Ex-ante
Foster innovation and employment	Society	Yes	Direct	Ex-post

Time savings				
Time savings	Data providers Data portals Data consumers	Yes	Direct	Ex-ante and Ex-post
Service quality improvement				
Higher satisfaction	Data consumers	No	Direct	Ex-post
Improve compliance	Data providers	Partially	Direct	Ex-post
Better data availability	Data consumers	No	Direct	Ex-ante and Ex-post
Spill-over effects				
Increase transparency	Society	No	Indirect	Ex-ante and Ex-post
Impact on growth and competitiveness	Society	Yes	Indirect	Ex-ante and Ex-post

6.2. Benefits for data consumers

The main direct benefits of using DCAT-AP for data consumers include: time savings, better data availability and higher service satisfaction.

The paper already presented how the adoption of common standards in Europe increases the discoverability of data. Due to the increased discoverability and better search functionalities powered by the common metadata specification, data consumers would spend less time finding, interpreting and processing data sets. DCAT-AP moreover improves the availability of data and metadata on the web, which involves making data available in usable formats.

6.3. Benefits for data providers and data portals

Benefits for actors on the supply side of data, i.e. data providers and data portals, include reduced operational costs, time savings, improved compliance and vendor lock-in avoidance.

DCAT-AP foster the automatic exchange and machine processing of metadata, which leads to significant time savings and operational cost savings for data providers and data portals. By complying with DCAT-AP and its extensions StatDCAT-AP and GeoDCAT-AP, data providers automatically comply with international standards and specifications such as DCAT, INSIRE, ISO 19115 and SDMX. By using open data standards and specifications in the design of a solution, the switching cost for the solution to change vendor or provider would significantly decrease, as data becomes portable to any other system that complies with the same standards. The use of DCAT-AP in the set-up of data portals make the data that resides in those systems portable to any other system that complies with DCAT-AP.

6.4. Benefits for society

A number of benefits from using DCAT-AP don't apply to one of the actors of the ecosystem, but more to society as a whole. These benefits mainly include spill-over effects from using interoperability solutions. In light of DCAT-AP, these include increased transparency, a positive impact on growth and competitiveness and fostered innovation and employment.

By increasing the discoverability of datasets, DCAT-AP and EDP support the objectives of public organisations to achieve their transparency goals. In the same line, better interoperability reduces barriers to cross-border business which has an impact on the competitiveness. DCAT-AP and the EDP support this benefit by significantly improving the cross-border availability and discoverability of datasets, which in its turn allows data consumers to develop innovative solutions based on the data.

7. FUTURE PLANS AND CONCLUSION

In this paper, we presented how Europe progresses steadily towards the creation of an open data ecosystem with the use of open standards built around the DCAT-AP specification.

DCAT-AP becomes currently a de facto open data standard in Europe with an increasing number of countries and portals adopting or extending it. As DCAT-AP is progressively rolled out by more portals, metadata will increase in quality, in turn making data more discoverable.

A number of next steps consist of:

a. Publishing more data: going beyond the low hanging fruit, to quote the European Data Portal's report on Open Data Maturity in Europe 2016[lxxiii].

b. Creating sustainable data infrastructures: Opening up data is not just about making sure vast amounts of data are published. It is also about ensuring sustainability. This means maintaining data platforms and access to data overtime.

c. Internationalise the discussion for the use of open standards for open government data: similar problems that have triggered the work around DCAT-AP in Europe exist in other countries and regions all over the world. Europe is actually paving the way for an international discussion to standardise open data descriptions.

d. There is still fragmentation in the open data ecosystem. Despite the efforts, different communities work separately and create "open data islands" that are not effectively connected. Examples include the geospatial, statistics and scientific & research data communities. Effort is still needed to create bridges and link these separate data worlds. Raising awareness and bringing all communities around the same table is necessary to avoid creating an open data babel tower in the future and paying later integration costs.

The ISA² Programme of the European Commission commits to maintaining the DCAT Application Profile, adapting to the needs of implementers and end-users while ensuring compliance with the DCAT specification. A new set of implementation guidelines is available for public review[lxxiv] and implementers can share their input for future revisions via the Joinup issue tracker[lxxv]. Last, the programme welcomes further suggestions and could be used as a vehicle for the facilitator role in the needed coordination across communities and domains.

End Notes

[i] Open Knowledge, 2015

[ii] European Commission, Creating Value Through Open Data, 2015.

https://www.europeandataportal.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf

[iii] Ibid

[iv] Ibid

[v] The International Open Data Charter, http://opendatacharter.net/

[vi] Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:en:PDF

[vii] 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/?uri=celex:32013L0037

[viii] European Commission, Commission Notice: 'Guidelines on recommended standard licences, datasets and charging for the re-use of documents', 2014

https://ec.europa.eu/digital-single-market/news/commission-notice-guidelines-recommended-standard-licences-datasets-and-charging-re-use

[ix] European Commission, Open Data Maturity in Europe 2016, 2016. URL:

https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n2_2016.pdf

[x] Ibid

[xi] Directive 2013/37/EU op.cit. .

[xii] European Commission, Open Data Maturity in Europe 2016, op.cit.

[xiii] Oude Luttighuis, Paul and Folmer, Erwin (2011) *Equipping the enterprise interoperability problem solver*. Interoperability in digital public services and administration: bridging E-Government and E-Business. Information Science Reference, pp. 339-354. ISBN 9781615208883

[xiv] http://www.aifb.kit.edu/images/a/a2/Iswc-semstats-2013-awa.pdf p.6

https://www.europeandataportal.eu/en/what-we-do

[xv]https://joinup.ec.europa.eu/sites/default/files/methodology_and_tools_for_metadata_governance_and_management_for_eu_institutions.pdf

[xvi] http://schema.org/

[xvii] https://www.w3.org/TR/void/

[xviii] http://ckan.org/

[xix] https://www.w3.org/TR/vocab-adms/

[xx] https://sdmx.org/

[xxi] http://inspire.ec.europa.eu/

[xxii] http://www.eurocris.org/cerif/main-features-cerif

[xxiii]https://joinup.ec.europa.eu/community/semic/news/isa%C2%B2-action-promoting-semantic-interoperability-amongst-european-union-member-s

[xxiv] F. Maali, R. Cyganiak, V. Peristeras, Enabling Interoperability of Government Data Catalogues, Lecture Notes in Computer Science, Vol. 6228, pp. 339-350, Springer, 2010

[xxv] W3C. Government Linked Data (GLD) Working Group. http://www.w3.org/2011/gld/wiki/Main_Page

[xxvi] W3C. Resource Description Framework (RDF). http://www.w3.org/RDF/

[xxvii] DCAT Application Profile for data portals in Europe Version 1.1, available on https://joinup.ec.europa.eu/catalogue/distribution/dcat-ap-version-11, p. 9.

[xxviii] W3C. The Organization Ontology. W3C Candidate Recommendation,25 June 2013.

http://www.w3.org/TR/2013/CR-vocab-org-20130625/

[xxix] IETF. BCP 47. Tags for Identifying Languages. http://www.rfc-editor.org/rfc/bcp/bcp47.txt

[xxx] D02.03.02: Updated specification DCAT-AP, pp. 4-5.

[xxxi] 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

[xxxii] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52014XC0724(01)

[xxxiii] European Commission. Joinup. DCAT application profile for data portals in Europe. GeoDCAT-AP working drafts. https://joinup.ec.europa.eu/node/139283

[xxxiv] https://joinup.ec.europa.eu/node/154143/

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

[xxxvi] https://joinup.ec.europa.eu/community/semic/og_page/core-vocabularies

[xxxvii] https://joinup.ec.europa.eu/community/semic/description

[xxxviii] https://www.govdata.de/standardisierung

[xxxix] https://joinup.ec.europa.eu/asset/dcat-ap_implementation_guidelines/asset_release/dcat-ap-how-extend-dcat-ap

[xl] https://www.mitre.org/publications/systems-engineering-guide/enterprise-engineering/engineering-informationintensive-enterprises/architectures-federation

[xli] https://www.europeandataportal.eu/en/what-we-do

[xlii] http://ec.europa.eu/transport/themes/infrastructure/ten-t-guidelines/project-funding/cef_en

[xliii] https://www.adaptlearning.org/

[xliv] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-mtec_en

[xlv] https://gitlab.com/european-data-portal/ckanext-edp/wikis/schema

[xlvi] https://joinup.ec.europa.eu/asset/dcat-ap_implementation_guidelines/issue/all

[xlvii] https://joinup.ec.europa.eu/asset/dcat-ap_implementation_guidelines/description

[xlviii] https://joinup.ec.europa.eu/node/150350/#AddTool

[xlix] ODS DCAT-AP Validator: https://joinup.ec.europa.eu/software/dcat-ap_validator/description

[l] https://validator.dcat-editor.com/

[li] http://geonetwork-opensource.org

[lii] http://entryscape.com

[liii] http://gptogc.esri.com

[liv] http://etl.linkedpipes.com/

[lv] https://github.com/opendata-swiss/ckanext-geocat

[lvi] https://github.com/ckan/ckanext-dcat

[lvii] https://github.com/Esri/geoportal-server/wiki/How-to-Publish-Resources#dcat

[lviii] https://github.com/ckan/ckanext-dcat

[lix] MDR: the metadata registry of the Publications Office of the EU: http://publications.europa.eu/mdr/

[lx] https://joinup.ec.europa.eu/asset/dcat-ap_implementation_guidelines/issue/mai1-mapping-national-themes-mdr-data-themes-vocabular

[lxi] https://joinup.ec.europa.eu/node/150359/

[lxii] https://joinup.ec.europa.eu/asset/dcat-ap_implementation_guidelines/issue/mi2-dataset-series

[lxiii] https://joinup.ec.europa.eu/asset/dcat-ap_implementation_guidelines/issue/mi3-provenance

[lxiv] https://joinup.ec.europa.eu/asset/dcat-ap_implementation_guidelines/issue/ui4-licence-documents-and-licence-uris

[lxv] D04.04 Implementation guidelines for the DCAT-AP, p. 18.

[lxvi] https://www.w3.org/TR/rdf11-concepts/#section-skolemization

[lxvii] https://joinup.ec.europa.eu/node/150348

[lxviii] E.g.https://joinup.ec.europa.eu/asset/dcat_application_profile/issue/mo12-grouping-datasets#comment-16648

[lxix] https://joinup.ec.europa.eu/asset/dcat-ap_implementation_guidelines/issue/mi2-dataset-series#comment-17725

[lxx] https://www.w3.org/TR/vocab-adms/#representation-technique

[lxxi] DG DIGIT/ISA Programme. Identification of IoP benefits (direct and indirect). Brussels, 2015.

[lxxii] D02.02 – Identification of IoP benefits (direct and indirect), ISA Programme, Luxembourg, 2016.

[lxxiii] Op.cit.

[lxxiv] https://joinup.ec.europa.eu/asset/dcat-ap_implementation_guidelines/asset_release/all

[lxxv] https://joinup.ec.europa.eu/asset/dcat_application_profile/issue/all

https://joinup.ec.europa.eu/asset/dcat_application_profile/issue/all