GeoDCAT-AP: Use cases and open issues

Authors Andrea Perego, Anders Friis-Christensen, Michael Lutz

Affiliation European Commission, Joint Research Centre (JRC) (https://ec.europa.eu/jrc/)

Abstract

This paper illustrates some issues and use cases identified during the design and implementation of GeoDCAT-AP, a metadata profile aiming to provide a representation of geospatial metadata compliant with the *DCAT application profile for European data portals* (DCAT-AP).

In particular, the paper focuses on those issues that may have a possible relevance also outside the geospatial domain, covering topics concerning metadata profile-based negotiation, publishing metadata on the Web, representing API-based data access in metadata, and approaches to modelling data quality.

Workshop *Smart Descriptions & Smarter Vocabularies* (SDSVoc). Amsterdam, 30 Nov - 1 Dec 2016 (https://www.w3.org/2016/11/sdsvoc/).

URL https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_25

Last modified 5 Dec 2016

Disclaimer: The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

Introduction

GeoDCAT-AP [8] is a metadata profile aiming to provide a representation of geospatial metadata compliant with the *DCAT application profile for European data portals* (DCAT-AP) [3].

DCAT-AP has been developed in the framework of the EU Programme *Interoperability solutions for public administrations, businesses and citizens* (ISA²) (http://ec.europa.eu/isa/) , with the purpose of defining a cross-domain metadata interchange format that can be used to share dataset metadata across data catalogues operated across the EU.

In this context, GeoDCAT-AP has been specifically designed to enable the sharing of geospatial metadata, in particular those available via the INSPIRE (http://inspire.ec.europa.eu/) infrastructure. To achieve this, GeoDCAT-AP defines mappings from ISO 19115 [12] (i.e., the world-wide standard for geospatial metadata) to DCAT-AP and other general-purpose RDF vocabularies.

During the design and implementation of GeoDCAT-AP, a number of issues have been identified, con-

cerning requirements for geospatial metadata that were not addressed in existing RDF vocabularies. The following sections focus in particular on those issues that may have a possible relevance also outside the geospatial domain. Each topic is introduced by first illustrating the problem statement. Then, the related use cases from GeoDCAT-AP are described, highlighting both the adopted solutions and the open issues.

Profile-based content negotiation

Problem statement

The ability to request metadata records following a given profile is a feature supported via query parameters by existing standards for catalogue services. E.g., this is the case for the Open Geospatial Consortium's *Catalogue Service for the Web* (CSW) [1] and the Open Archive Initiative's *Protocol for Metadata Harvesting* (OAI-PMH) [17].

The issue is that in all these cases the mechanism to request the output profile is specific to the interface used. An interface-independent mechanism, e.g., based on the use of specific HTTP headers, would be desirable for interoperability purposes.

Use case

The work concerning the implementation of GeoDCAT-AP included the development of an API, able to transform on the fly ISO 19139 records (ISO 19139 [13] defines the standard XML-based implementation of ISO 19115).

The GeoDCAT-AP API [22] is able to return records both in the DCAT-AP and GeoDCAT-AP profiles. The former is meant to be used for harvesting purposes by catalogues only supporting the basic DCAT-AP profile, whereas the latter can be used when a catalogue is supporting also the additional metadata elements defined in GeoDCAT-AP.

More precisely, the API has been designed to re-use the standard query interface of geospatial catalogue services (i.e., CSWs [1]), which already supports the possibility of specifying the output metadata profile and format. The output format can also be omitted: in such a case, the GeoDCAT-AP API determines it based on HTTP content negotiation. This is however not the case for the output metadata profile, since there is currently no HTTP-based mechanism for specifying this type of requests. So, the default output metadata profile is determined by the API settings.

The approach currently used in the GeoDCAT-AP API to address this issue is to return metadata records along with a set of HTTP Link headers [19] specifying:

- The URL of the source ISO 19139 record.
- The metadata profile of the returned record.
- The URLs of the alternative RDF serialisations of the returned record.

A possibility is to modify the last set of Link headers to include also the URLs of the alternative profiles of the returned record. In this scenario, getting access to the preferred metadata profile and serialisation would require retrieving all the different alternative representations until the relevant one is found.

Publishing metadata on the Web

Problem statement

Geospatial data catalogues offer effective discovery functionalities for specialists. However, as in most data portals, using basic free text search is usually far from being a satisfactory experience. Actually, users (both non-experts and specialists), when searching for data, are frequently making use of popular search engines as a first step to get to the data they are looking for. The same can be said for any other domain-specific and general-purpose data catalogue.

Improving free text search in data catalogues is unlikely to address this issue. Moreover, it would not help users who do not know in which catalogue(s) the relevant data are available. Users will keep on using in any case search engines for this purpose.

An option would be optimising data catalogues for Web discovery, by implementing consistently SEO (Search Engine Optimisation) techniques. The advantages include (but are not limited to) the following:

- Increase data visibility - as well as of the catalogues giving access to them.
- Enabling users to find the data better satisfying their requirements (in terms of spatial and temporal coverage, granularity, formats, access and licensing conditions, etc.).
- (For geospatial metadata) Enabling queries not limited to free text fields (e.g., dataset title and description), but concerning also coordinates, spatial / temporal resolution, etc.

Use case

The output formats supported by the GeoDCAT-AP API [22] include all the existing RDF serialisations, including HTML+RDFa.

HTML+RDFa [11] is W3C standard defining a mechanism to embed structured metadata in HTML documents. HTML+RDFa is actually one of the mechanisms used by search engines to index Web pages.

The GeoDCAT-AP API is not the only tool addressing this issue. For example, this is the case of one of the prototypes developed in the framework of the Geonovum testbed [20], where an extension to the GeoNetwork platform [9] has been developed, able to transform ISO 19139 records into HTML+RDFa.

Additional work in this direction concerns the use of the Schema.org vocabulary to annotate Web pages via HTML+RDFa, or other mechanisms - as Microformats [15], Microdata [10] and HTML-embedded JSON-LD [14] - see Section 6.20 (https://www.w3.org/TR/json-ld/#embedding-json-ld-in-html-documents).

Schema.org (http://schema.org) is an initiative backed up by companies running popular search engines (Google, Microsoft, Yahoo and Yandex) to develop a general-purpose vocabulary for the annotation of Web pages. As such, its scope is not limited to data, but it includes also domains as e-commerce, businesses, audio-visual content.

Some work has been carried out to define mappings for DCAT-AP [5], whereas the Geonovum testbed mentioned earlier defined mappings for ISO 19115. In both cases, a number of gaps have been identified, making Schema.org not able to provide a full mapping of DCAT-AP and ISO 19115. Such gaps do not concern only domain-specific metadata elements (e.g., spatial resolution, coordinate reference system), but also more general requirements, as the modelling of identifiers and terms from tax-

onomies, thesauri, classification systems.

On this topic, it is worth noting that Google recently started a project, titled *Science Datasets* (https://developers.google.com/search/docs/data-types/datasets), aiming to investigate how to improve the description and discovery of scientific datasets by using Schema.org. On the other hand, the mapping exercises mentioned above raise also the question on whether a full mapping is really needed, given that the mapped vocabularies address different use cases.

Modelling service/API-based data access

Problem statement

This is a domain-independent issue, but a key one for geospatial metadata.

In the geospatial domain, data are typically made accessible via services (e.g., a view or download service), that, to be used, require specific clients. In metadata, the link to such services is usually pointing to an XML document describing the service's "capabilities". This of course puzzles non-expert users, who expect instead to get the actual "data".

Moreover, an additional issue is that a service may provide access to more than one dataset. As a consequence, users do not know how to get access to the subset of relevant data accessible via a service.

Some catalogue platforms (as GeoNetwork [9] and, to some extent, CKAN [2]) are able to make this transparent for some services (typically, view services), but not for all. It would therefore be desirable to agree on a cross-domain and cross-platform approach to deal with this issue.

Use case

In the framework of the *DCAT-AP implementation guidelines* [4], a proposal has been developed to describe a service/API by using an OpenSearch document [16] - see issue *DT2: Service-based data access* (https://joinup.ec.europa.eu/asset/dcat_application_profile/issue/dt2-service-based-data-access) on JoinUp.

This approach has the advantage of adopting a general-purpose solution (applicable also to non-geo services), that could be used by catalogues as well as Web applications to make data access more transparent to users, or at least to provide some guidance on how to use the service. Moreover, the same approach would enable machine-based data access.

Modelling data quality

Problem statement

Here, the notion of "data quality" is used in its broadest sense, including:

- Fit-for-purpose.
- Data precision / accuracy.
- Compliance with given quality benchmarks, standards, specifications.
- Quality assessments based on data review / users' feedback.

There is of course no single solution to address all this, so, the following use cases focus on some practical issues identified during the development of DCAT-AP and its domain-specific extensions - namely, GeoDCAT-AP (for geospatial metadata) and StatDCAT-AP (for statistical metadata) [21].

Use case

Currently, the DCAT-AP family of metadata profiles cover the following aspect of data quality:

- Metadata conformance with a metadata standard (DCAT-AP).
- Data conformance with a given data schema/model (DCAT-AP).
- Data conformance with a given reference system (spatial or temporal) (GeoDCAT-AP).
- Data conformance with a given quality specification / benchmark (GeoDCAT-AP).
- Associating data with a quality report (StatDCAT-AP).

The first four ones are modelled by using `dct:conformsTo` (http://purl.org/dc/terms/#terms-conformsTo), whereas the last one makes use of the W3C Data Quality Vocabulary (DQV) [6]. Moreover, GeoDCAT-AP provides an alternative and extended way of expressing "conformance" by using the W3C PROV Ontology [18], allowing the specification of additional information about conformance tests (when this has been carried out, by whom, etc.).

An example of the GeoDCAT-AP PROV-based representation of conformance is provided by the following code snippet:

EXAMPLE

```
a:Dataset a dcat:Dataset ;
  prov:wasUsedBy a:TestingActivity .

a:TestingActivity a prov:Activity ;
  prov:generated a:TestResult ;
  prov:qualifiedAssociation [ a prov:Association ;
# Here you can specify which is the agent who did the test, when, etc.
    prov:hadPlan a:ConformanceTest ] .

# Conformance test result
a:TestResult a prov:Entity ;
  dct:type <http://inspire.ec.europa.eu/metadata-codelist/DegreeOfConformity/co
nformant> .

a:ConformanceTest a prov:Plan ;
# Here you can specify additional information on the test
  prov:wasDerivedFrom <http://data.europa.eu/eli/reg/2014/1312/oj> .

# Reference standard / specification
<http://data.europa.eu/eli/reg/2014/1312/oj> a prov:Entity, dct:Standard ;
  dct:title "Commission Regulation (EU) No 1089/2010 of 23 November 2010 implem
enting Directive 2007/2/EC of the European Parliament and of the Council as reg
ards interoperability of spatial data sets and services"@en
  dct:issued "2010-11-23"^^xsd:date .
```

An example of the GeoDCAT-AP PROV-based representation of conformance. The key entities are marked in bold.

There are however some aspects of data quality not yet addressed. As far as GeoDCAT-AP is concerned, these include the following ones:

- Spatial resolution.
- Data quality assessments expressed with quantitative test results.

Finally, another potentially relevant use case is related to exploiting users' feedback to assess data quality. This may include feedback from data consumers (e.g., reporting errors in a dataset), as well as data peer reviewing.

This information need not be part of dataset metadata, and it can be collected and stored separately. Requirements include contextual information (who's providing feedback, when, etc.), that can be used to aggregate feedback data according to given criteria.

To address such gaps, the possible use of DQV and of the W3C Dataset Usage Vocabulary (DUV) [7] is being considered. In particular:

- The DQV specification includes a number of examples describing how to model spatial resolution - see Section 6.13 (https://www.w3.org/TR/vocab-dqv/#ExpressDatasetAccuracyPrecision). Notably, such an approach can also be used to model also "temporal resolution", and, more in general, the notion of "data granularity".
- DQV also provides a means to express quantitative quality assessments - see Section 6.1 (https://www.w3.org/TR/vocab-dqv/#ExpressAQualityAssessmentWithQualityMetrics).

When used in conjunction with DQV, DUV allows the representation of annotations about data quality, carrying also provenance information.

References

[1] *Catalogue Service* (2016) Open Geospatial Consortium (OGC) http://www.opengeospatial.org/standards/cat

[2] *ckan - The open source data portal software* (2016) Open Knowledge Foundation (OKFN) http://ckan.org/

[3] *DCAT application profile for data portals in Europe* (2015) EU ISA Programme (ISA²) https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe

[4] *DCAT application profile implementation guidelines* (2016) EU ISA Programme (ISA²) https://joinup.ec.europa.eu/solution/dcat-application-profile-implementation-guidelines

[5] *DCAT-AP to Schema.org Mapping* (2016) European Commission, Joint Research Centre (JRC) https://webgate.ec.europa.eu/CITnet/stash/projects/ODCKAN/repos/dcat-ap-to-schema.org/

[6] *Data on the Web Best Practices: Data Quality Vocabulary* (2016) World Wide Web Consortium (W3C) https://www.w3.org/TR/vocab-dqv/

[7] *Data on the Web Best Practices: Dataset Usage Vocabulary* (2016) World Wide Web Consortium (W3C) https://www.w3.org/TR/vocab-duv/

[8] *GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe* (2016) EU ISA Programme (ISA²) https://joinup.ec.europa.eu/node/139283

[9] *GeoNetwork opensource* (2016) Open Source Geospatial Foundation (OSGeo) http://geonetwork-opensource.org/

[10] *HTML Microdata* (2013) World Wide Web Consortium (W3C) https://www.w3.org/TR/microdata/

[11] *HTML+RDFa 1.1 - Second Edition: Support for RDFa in HTML4 and HTML5* (2015) World Wide Web Consortium (W3C) https://www.w3.org/TR/html-rdfa/

[12] *ISO 19115:2003: Geographic information -- Metadata* (2003) International Organization for Standardization (ISO) https://www.iso.org/standard/26020.html

[13] *ISO/TS 19139:2007: Geographic information -- Metadata -- XML schema implementation* (2007) International Organization for Standardization (ISO) https://www.iso.org/standard/32557.html

[14] *JSON-LD 1.0: A JSON-based Serialization for Linked Data* (2014) World Wide Web Consortium (W3C) https://www.w3.org/TR/json-ld/

[15] *Microformats* (2016) http://microformats.org/about

[16] *OpenSearch* (2016) OpenSearch.org http://www.opensearch.org/

[17] *Protocol for Metadata Harvesting* (2002) Open Archives Initiative (OAI) https://www.openarchives.org/pmh/

[18] *PROV-O: The PROV Ontology* (2013) World Wide Web Consortium (W3C) https://www.w3.org/TR/prov-o/

[19] *RFC 5988: Web Linking* (2010) Internet Engineering Task Force (IETF) https://tools.ietf.org/html/rfc5988

[20] *Spatial Data on the Web using the current SDI* (2016) Geonovum https://geo4web-testbed.github.io/topic4/

[21] *StatDCAT application profile for data portals in Europe* (2016) EU ISA Programme (ISA²) https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/

[22] *The GeoDCAT-AP API (GeoDCAT-API)* (2016) EU ISA Programme (ISA²) https://webgate.ec.europa.eu/CITnet/stash/projects/ODCKAN/repos/iso-19139-to-dcat-ap/browse/api