Applying DCAT vocabulary on RDF datasets

Ghislain A. Atemezing
Mondeca, 35, bd de Strasbourg, 75010 Paris, France
Email: ghislain.atemezing@mondeca.com

Abstract

In this paper, we show the implementation of DCAT on geospatial data published in RDF for ING France. We consider an endpoint as a catalogue of RDF datasets, where each named graphs can be versioned and thus can be implemented. We highlight the choices made to manage the versions of the data often updated in a public endpoint, without breaking consumption by users of the data.

Keywords: DCAT implementation, geodata, data.ign.fr, versioning, RDF management

Introduction

Since 2012, the endpoint at http://data.ign.fr/id/sparql offers to users RDF datasets for geospatial data as part of the agency to make publish in structured from open data related to French territories. However, after the first release of the data, the new challenges are to enhance data discovery and manage the versions of the dataset in the RDF store. We present here the current approach to attach to each dataset in the endpoint represented by a different named graph using a mix of available vocabularies: DCAT[1], PAV[2], SPARQL-SD[3] and DCTERMS[4]. The goal is to converge towards the best approach to provide machine readable metadata on top of endpoints, different from statistical information. The remainder of the paper is as follows. We present the motivations and the approach implemented, then how we model versioning between datasets. Finally, we conclude by a brief summary.

Metadata Implementation Approach

A catalog is a collection of RDF datasets represented by one or many Named Graph (NdG). Ideally there may be many catalogs in an endpoint, but for simplicity we constraint that an endpoint contains at least one catalog, defined by the class `dcat:Catalog`. The catalog is linked to all the different datasets (including all the versions) available for retrieval in the

[1] https://www.w3.org/TR/vocab-dcat/
[2] http://purl.org/pav/
[3] https://www.w3.org/ns/sparql-service-description
[4] http://purl.org/dc/terms

endpoint by the property `dcat:dataset`. The URI used to create such resource could be of the form http://{domain}/id/set/catalog/{id}. E.g., <http://data.ign.fr/id/set/catalog/1>

This sample below (Listing 1) the describes an excerpt of a catalogue for the RDF datasets in the data.ign.fr endpoint.

```
<http://data.ign.fr/id/set/catalog/1> a dcat:Catalog> ;
       <http://purl.org/dc/terms/issued> "2016-05-25"^^xsd:date ;
       dc:title "Catalog of IGN France's datasets published consistently with
semantic Web standards as part of the research project Datalift."@en ;
       dc:publisher
<http://fr.dbpedia.org/resource/Institut_national_de_l%27information_g%C3%A9ograp
hique_et_foresti%C3%A8re> ;
       dc:rights "Copyright 2016, IGN" ;
       dc:language
<http://id.loc.gov/vocabulary/iso639-1/fr>,<http://id.loc.gov/vocabulary/iso639-1
/en> ;
       dcat:dataset              <http://data.ign.fr/id/set/dataset/ignf/20140409>
,<http://data.ign.fr/id/set/dataset/ignf>,<http://data.ign.fr/id/set/dataset/geof
la> .
```

Listing 1: There are three datasets in the catalogue, with two versions of one dataset.

The URI for each named graph can be represented in the form: http://{domain}/id/{dataset_name}/namedgraph}, while the URI pattern for each dataset is represented as http://{domain}/id/set/dataset/{dataset_name}. Example: http://data.ign.fr/id/set/dataset/geofla for the dataset on administrative boundaries of Metropolitan France. Figure 1 depicts the model used to implement an endpoint as a collection of RDF datasets and relationships with various DCAT classes.

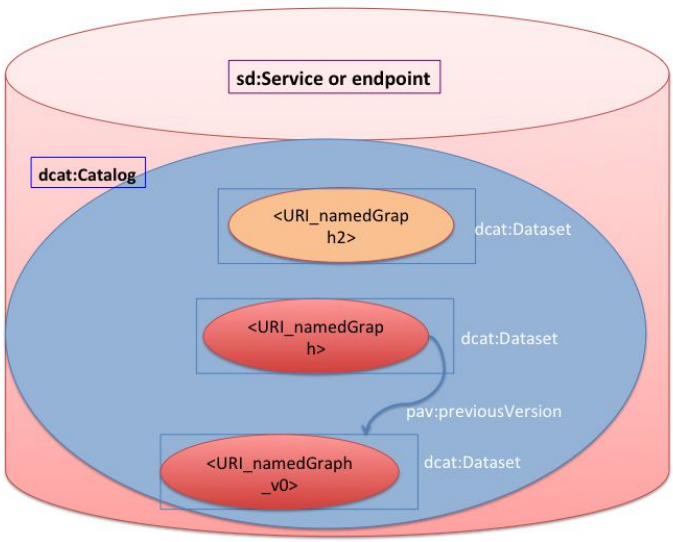

Figure 1: Catalog of datasets with named graphs in an endpoint

A dataset might contains one or many named graphs linked from a dcat:Dataset by the property `sd:namedGraph`. The distribution of the RDF dataset published in an endpoint (without a dump) is limited to at most one, providing the `dcat:accessURL` to be the endpoint URI and the media types limited to the formats handled by "SELECT" or ASK queries. We consider the CONSTRUCT queries as a way to create a new "dataset", which is out of scope of this work. By this means, we can add the number of triples in each named graph by using the property `void:triples`. Listing 2 shows how this is implemented for the geofla dataset.

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix sd: <http://www.w3.org/ns/sparql-service-description#> .

<http://data.ign.fr/id/set/dataset/geofla>
  a dcat:Dataset ;
    sd:namedGraph _:g1, _:g2, _:g3, _:g4, _:g5 ;
  dcat:distribution [ a dcat:Distribution;
                      dc:title "Most recent version of the dataset on
administrative boundaries of Metropolitan France";
                      dcat:accessURL <http://data.ign.fr/id/sparql> ;
                      dcat:mediaType "text/csv", "application/rdf+xml","text/html"
                     ] .


_:g1 a sd:NamedGraph ;
       sd:name <http://data.ign.fr/id/geofla/commune/> ;
       sd:graph [a sd:Graph ; dct:title "communes from IGN"; void:triples
"1061267"^^xsd:decimal ] .

_:g2 a sd:NamedGraph ;
       sd:name <http://data.ign.fr/id/geofla/canton/> ;
       sd:graph [a sd:Graph ; dct:title "Cantons Geofal"; void:triples
"1658342"^^xsd:decimal ] .


_:g3 a sd:NamedGraph ;
       sd:name <http://data.ign.fr/id/geofla/arrondissement/> ;
       sd:graph [a sd:Graph ; dct:title "Arrondissements from IGN"; void:triples
"7245"^^xsd:decimal ] .

_:g4 a sd:NamedGraph ;
       sd:name <http://data.ign.fr/id/geofla/departement/> ;
       sd:graph [a sd:Graph ; dct:title "Departments from IGN"; void:triples
"405562"^^xsd:decimal ] .
```

```
_:g5 a sd:NamedGraph ;
       sd:name <http://data.ign.fr/id/geofla/region/> ;
       sd:graph [a sd:Graph ; dct:title "Regions from IGN"; void:triples
"428305"^^xsd:decimal ] .
```

Listing 2: An excerpt of a dataset with the distribution and corresponding named graphs.

Dataset versioning Management

We assume that each dataset can evolve and to create the URI for them, we use the modified date to generate the "snapshot" of the given dataset. Thus at any moment, the current version of the data is the URI created at the initialization of the catalogue. We use the `pav:previousVersion` to link different versions of the same dataset. Listing 3 describes a dataset with one prior version.

In the case of a single named graph, the pattern for the named graph is http://{domain}/id/{dataset_name}/ as depicted in Listing 3. This choice

```
<http://data.ign.fr/id/set/dataset/ignf> a dcat:Dataset ;
       dc:issued "2016-05-25"^^<http://www.w3.org/2001/XMLSchema#date> ;
       dc:title  "IGN France's registry of coordinate reference systems version
2.1.3"@en ;
       dc:publisher
<http://fr.dbpedia.org/resource/Institut_national_de_l%27information_g%C3%A9ograp
hique_et_foresti%C3%A8re> ;
      sd:namedGraph [ sd:name <http://data.ign.fr/id/ignf/> ;
                        sd:graph [ dct:title "IGNF registry of coordinate reference
systems"@en; void:triples "24062"^^xsd:decimal] ;
       dcat:distribution [ dc:title "Most recent version of the registry of
coordinate reference systems accessible through a SPARQL endpoint"@en ;
                           dcat:accessURL <http://data.ign.fr/id/sparql> ;
                           dcat:mediaType "text/csv",
"application/rdf+xml","text/html" ;
                          dc:rights "Copyright 2016, IGN"
                          ];
       dc:language <http://id.loc.gov/vocabulary/iso639-1/fr> ;
       dcat:keyword "coordinate reference systems"@en , "geodetic features"@en ;
       dcat:landingPage <http://data.ign.fr/data.html> ;
       dcat:contactPoint <http://data.ign.fr/contacts> ;
       dc:spatial <http://fr.dbpedia.org/resource/France> ;
       pav:previousVersion <http://data.ign.fr/id/set/dataset/ignf/20140409>.
```

Listing 3: Versioning sample in the case of the IGNF dataset, when a dataset has a single named graph

Conclusion

This work presented an implementation of RDF endpoint using DCAT. Based on the choices to represent a catalogue, we can infer that a `sd:GraphCollection` represents a `sd:Dataset`. One benefit of this implementation is that it makes it easier to access the current version of a dataset using SPARQL query. The current dataset is linked to its "snapshots" using `pav:previousVersion`. The URI used to link the version returns the metadata of the version with the distribution URI (an HTML page when dereferenced), where there is a link to dereferenced the whole dataset (a trig file). A future work is to find out how to embed the metadata of this version in the HTML page itself. The approach presented in this paper is currently implemented in a public endpoint. However, what could be interesting is to see how consumers discover easily datasets based on the proposed approach. Also, it could be benefits to compare our approach to other publishers of geospatial data on the web.