An Abstract Specification Technique for the Publication of Linked Geospatial Data

Gilles Falquet, Claudine Métral
Université de Genève
Centre universitaire d'informatique
7 route de Drize
1227 Carouge, Switzerland
claudine.metral | gilles.falquet@unige.ch

Sylvain Ozaine, Gregory Giuliani
Université de Genève,
Institut des sciences de l'environnement
66 bd Carl-Vogt
1205 Genève, Switzerland
s.o@vs.ch, gregory.giuliani@unige.ch

Abstract

The publication of linked geodata on the semantic web may involve relatively complex operations to map source data to their published version and to link these ones to other datasets. In this paper we propose a technique to provide a compact and abstract description of this process. This technique is based on the use of RDF graph mapping rules that are expressed in the SPARQL Semantic web query language. We show that such rules are sufficient to express complex mapping and linking operations at a high level, without referring to operations specific to a particular vendor or tool. We also show how this technique applies to a non-trivial use case in the domain of tropical cyclones.
Keywords: Linked geodata, Semantic web, RDF, SPARQL, Publication metadata

1 Introduction and background

The Semantic web is becoming a convenient medium to make data and in particular geo-referenced data available to a wide range of applications. The publication of data on the Semantic web essentially requires two operations: transforming the data into an RDF graph, i.e. a set of (*subject*, *predicate*, *object*) triples, and creating links between this graph and other, published, graphs.

Several methodologies or methodological elements have been developed to help carrying out the Semantic web publication process. But most existing methodological guidelines don't propose a way to formally document the publishing workflow. As a consequence important provenance information are lost, and the process is difficult to repeat.

We propose a metadata model to represent (1) the goal and requirement analysis that lead to the publication of data sources, (2) the mappings that produced the published data from the available sources; (3) how entities from different sources were matched and linked. This model relies on an ontology of geodata operations. In this paper we present the developed model and a use case on which the model has been tested.

Giving a meaning to RDF triples by accurately defining predicates and objects requires the use of vocabularies and ontologies. An ontology offers a formal description of a given knowledge domain including the definition of its classes, types, properties and hierarchy. While the RDF Schema recommendation provides a data-modelling vocabulary for RDF data[1], the Web Ontology Language (OWL), an extension of RDF Schema, offers a formal syntax for writing Web ontologies[2].

Among vocabularies and ontologies describing geo-concepts and publicly available, one can cite the FAO geopolitical ontology[3],the SWEET ontologies[4] providing an earth and environmental terminology or the geospatial foundation ontologies[5], representing geospatial concepts and properties for use on the Worldwide Web.

Datasets can be stored in RDF stores, also known as triplestores (Heath & Bizer 2011). Triplestores can be queried using the SPARQL query language (Garlik & Seaborne 2013) through SPARQL endpoints. SPARQL allows exploring a set of RDF structured data by interrogating but also adding, modifying or deleting data in a triplestore. SPARQL also enables building new RDF graphs using CONSTRUCT queries. GeoSPARQL[6] defines a vocabulary for representing geospatial data in RDF as well as an extension to the SPARQL query language for processing geodata.

A few methodologies have been proposed for the publication of data or geo-data on the Semantic web. For instance Vilches-Blázquez et al. (2014) propose a publication process that comprises the following steps:

- specification of requirements (identification and analysis of geospatial data sources, URI design)

[1] https://www.w3.org/TR/rdf-schema/
[2] http://www.w3.org/OWL/
[3] http://www.fao.org/countryprofiles/geoinfo/geopolitical/resource/geopolitical.org
[4] https://sweet.jpl.nasa.gov/
[5] https://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/
[6] http://www.opengeospatial.org/standards/geosparql

- ontology modelling (in order to determine the ontology that will be used for representing the data sources)
- RDF generation
- links generation to other datasets
- data and metadata publication and exploitation

In many cases publishing a dataset on the semantic web does not reduce to directly translating from an existing data model (in general the relational data model) to RDF. The published data are in fact a view on the source data. This view must satisfy the publication requirements set by the publisher/data owner in terms of:

- structural complexity: the published schema must be sufficiently simple to facilitate the development of applications;
- data complexity (levels of detail): the desired level of detail of the published data can differ from the level of detail of the source data;
- pre-processing: some complex computations may be carried out once for all to relieve the application developers from the (re-)implementation of these computations;
- coverage/completeness: the publication may require the inference of missing data;
- confidentiality and anonymization: only a part of the available data may be released to the public;
- etc.

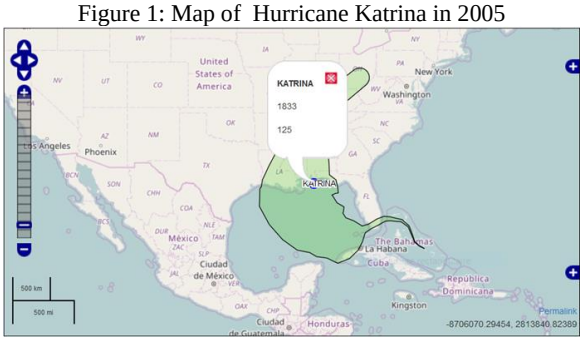
In other words, the published data must be usable for a set of intended applications. Which of course does not prevent their use for the creation of other, unforeseen, applications.

We tested the proposed publication metadata model on a use case related to tropical cyclones. There exists currently two data sources that are major resources: (1) the PREVIEW Global Risk Data Platform[7] which aims providing spatial data on global risks linked with various natural hazards such as tropical cyclones and associated storm surges, droughts, earthquakes, forest fires, floods, landslides, tsunamis and volcanic eruptions (Giuliani & Peduzzi 2011); (2) The EM-DAT international disaster database (Emergency Events Database: http://www.emdat.be) that provides access to data on more than 22,000 major disasters worlwide from 1900 to the present day. Although this database offers access to many statistics on past human and financial losses, EM-DAT events and figures are not precisely georeferenced as their only available spatial information are administrative boundaries. These two data sources are complementary and could benefit from being interconnected.

The PREVIEW dataset that we used includes the evolution of the spatial extent of tropical cyclones events between 1970 and 2014, modelled as polygon buffers from the tracks of the IBTrACS database[8] using a formula taking into account central pressure, wind speed and other variables (Peduzzi et al. 2005). Every distinct event is associated to one or several buffers describing the different steps of its evolution. Every buffer is itself associated to the surface of an impacted country or to a sea surface, and also to a Saffir-Simpson category. We also used an EM-DAT database subset including disaster figures (such as the number of affected people or the estimated damages) on tropical cyclones from 1970 to 2011. In this database disaster figures are linked with one event and one impacted country.

The purpose, among other things, is to publish cyclone data as maps such as illustrated in figure 1.

Figure 1: Map of Hurricane Katrina in 2005



2 Abstract specification of the publishing transformation

The basic idea is to specify the publication process as a graph mapping from the source data graph to the publication data graph.

The source data graph is an RDF graph obtained by representing every data source as a graph (using standard transformation from the relational or other models to RDF) and taking the union of these graphs.

The publication graph is the union of the subgraphs obtained by applying transformation rules on subgraphs of the source graph. Although several graph transformation formalisms can be found in the literature (Ehrig & al., 2006), we have remarked that the SPARQL query language for RDF, with a few extensions, provides a very convenient and readable alternative to these formalisms. Therefore the graph mapping will be specified by a set of (extended) SPARQL CONSTRUCT queries that construct graph elements (nodes and edges) from selected subgraphs that correspond to graph patterns. The general form of such queries is therefore:

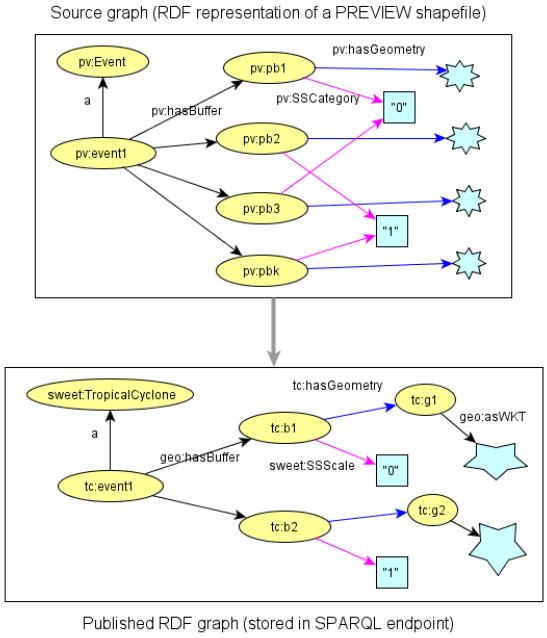
construct
publication graph triples
where
source graph selection patterns
specification (bindings) of the publication triple variables

In our use case, as described in section 1, source data from PREVIEW describe events. Each event is associated to one or several buffers. Each buffer is itself associated to a surface (polygon) and to a Saffir-Simpson category. In the publication graph we want to associate each tropical cyclone (a PREVIEW event) to a buffer that describes the affected countries for a given Saffir-Simpson category. Thus a buffer must be associated to a geometry (a multipolygon) and a category. We also want to represent the geometries as WKT literals. To obtain such a WKT geometry representation, we have to simplify the downloaded PREVIEW shapefile. This

[7] http://preview.grid.unep.ch
[8] http://www.ncdc.noaa.gov/ibtracs/

step is mandatory since many events of the original dataset are associated to country-clipped polygons that can be complex and have a very high number of vertices. We also have to merge the polygons that are associated to the same category. We thus obtain multipolygons, each of which being associated to a Saffir-Simpson category. We also have to apply a conversion of the geometry format into a string WKT representation. This can be graphically represented as in figure 2.

Figure 2: Source and published RDF graphs



This transformation can be specified with the following CONSTRUCT query:

```
prefix tc:  <http://geolink.grid.unep.ch/tropcyc#>
prefix pv:  <http://geolink.grid.unep.ch/preview#>
prefix geo: <http://www.opengis.net/ont/geosparql#>
prefix sweet:
   <http://sweet.jpl.nasa.gov/2.3/phenAtmoPressure.owl#">
construct {
 ?c a sweet:TropicalCyclone; tc:hasBuffer ?b.
 ?b sweet:SSScale ?cat; geo:hasGeometry ?g.
 ?g geo:asWKT ?w.
}
where {
  ?e a pv:Event;
     pv:hasBuffer ?pb.
   ?pb pv:SSCategory ?cat;
       pv:hasGeometry ?pg.
   # publication triple variables
   bind(genCycIRI(?e) as ?c)
   bind(genBufIRI(?e, ?cat) as ?b)
   bind(genGeoIRI(?e, ?cat) as ?g)
   bind(toWKT(Simplify(Merge(?pg))) as ?w)
}
group by ?cat
```

This specification is abstract because (1) the geometric functions appearing in the bindings, such as toWKT, Simplify, or Merge, refer to classes of operations (defined in an ontology of spatial operations) not to a specific implementation, (2) some functions are left unspecified (e.g. the IRI generating functions genCycIRI, genBufIRI, and genGeoIRI.

The IRI generating functions play an important role because they ensure that the same entity is represented by the same IRI in all the subgraphs generated by different SPARQL CONSTRUCTs (mapping rules). For instance the binding:

```
bind(genBufIRI(?e, ?cat) as ?b)
```

means that the IRI of a buffer in the publication graph is obtained by combining the IRI of the event and the IRI of the Saffir-Simpson category. IRI generation functions must be injective, i.e, they must not generate the same IRI for different parameters.

In addition to transformation rules that take as input the source graph and produce parts of the publication graph, some rules are intended to augment the publication graph. These rules have as input the publication graph and other data sources, and they produce new triples to be added the publication graph.

For instance, in the running example the above-described rule produced new polygons for each cyclone. Then the name of the affected countries was attributed to each resulting polygon using the "TM World Borders" shapefile retrieved from the thematicmapping.org website and performing ArcGIS Spatial Join-One-To-Many operations.

This can be specifed by the following CONSTRUCT query (on an RDF translation of the World Borders Dataset):

```
prefix tmwb:
<http://thematicmapping.org/downloads/world_borders.php#>
prefix fao: <http://fao.270a.info/dataset/>
prefix tc:  <http://geolink.grid.unep.ch/tropcyc#>
prefix geo: <http://www.opengis.net/ont/geosparql#>

construct {
 ?b tc:country [fao:codeISO3 ?cISO3;
                fao:nameListEN ?cname]
}
where {
 ?b geo:hasGeometry/geo:asWKT ?bPolygon.
 ?cb a tmwb:CountryBoundary;
     tmwb:shape ?cPolygon;
     tmwb:name ?cname
     tmwb:ISO3 ?cISO3.
 filter(intersects(?bPolygon, ?cPolygon)
}
```

3 Describing the entity linking and matching operations

If the published data come from more than one source, there is necessarily at least one type of entities that must be matched or linked between the two sources (otherwise the sources could be published separately). If the published data are

linked to other data on the semantic web there has been a matching or linking process to interconnect these data sets.

By matching process we understand a process that finds pairs of entities (represented by RDF nodes) that represent the same real-world entities. Matching algorithms can range from simple attribute value comparison to complex geometric operations (when working with different degrees of accuracies or different dimensionalities) or to machine learning techniques.

By linking process we understand a process that finds pairs of entities that are semantically related (through a relationship of interest for the data publication // a relationship present in the publication schema). At the RDF level we use the built-in OWL property owl:sameAs to indicate that two URI references actually refer to the same entity.

Again, this kind of linking rule can be abstractly specified as a CONSTRUCT operation of the form:

```
construct {?o1 owl:sameAs ?o2}

where {
  selection condition on ?o1
  selection condition on ?o2
  filter(MatchingFunction(?o1, ?o2, other parameters)
}
```

where the matching function can be based on complex algorithms and/or human processing.

In the use case, since one of the goals was to publish the PREVIEW cyclone information enriched with data from EMDAT, it was necessary to match the cyclone entities of the two sources. In this case the matching process was based on the name, year, start month and affected country properties and included a human cleaning of the event names, which leads to the following specification

```
prefix em: <http://geolink.grid.unep.ch/emdat#>
prefix pv:  <http://geolink.grid.unep.ch/preview#>

construct {?emCyc owl:sameAs ?prevCyc}

where {?emCyc a em:Cyclone; em:name ?ecName; ... .
   ?prevCyc a pv:Cyclone; pv:name ?pcName; ... .
  filter(MatchingFunction(?ecName, ?pcName, ... ))
}
```

Despite the prior editing of cyclone names, this operation pointed to some important discrepancies between the two datasets. Besides unresolved nomenclature problems and mismatched start dates of events, we observed that the complete lists of cyclones differ, as some events from the PREVIEW dataset are missing in the EM-DAT one and vice versa. Therefore, the specification of the matching function is a combination of automated processing (exact name and data matching) and human processing (matching by human inspection).

We also had to realize a similar linking operation to interlink the PREVIEW events with corresponding entries in DBpedia. The main purpose of this linking was to obtain additional information, such as the number of inhabitants in the impacted cities.

4 Proposed publication metadata

The specification of the publishing transformation and of the entity matching and linking can be made available to the data users by publishing them as metadata. In their simplest form these metadata are strings representing the SPARQL queries that make up the specifications. If needed, the queries can be represented as RDF triples by using the scheme proposed by Knublauch (2013).

To make these specifications understandable, they must be accompanied by the RDF schema(s) of the data source(s) and by the publication schema. Therefore the metadata associated with a geographic dataset published on the semantic Web must include :sourceVocabulary and publicationVocabulary properties. For the running example it takes the following form[9]:

```
:CycloneInfo a :GeodataPublication ;
  :sourceVocabulary <http://geolink.grid.unep.ch/emdat> ;
  :sourceVocabulary <http://geolink.grid.unep.ch/preview> ;
  :publicationVocabulary <http://geolink.grid.unep.ch/tropcyc> ;

  :mappingRule [ :name "Generate simplified polygons" ;
    :expression "prefix .... construct ..." ;
    :spatialOperation [a oso:Simplify ;
                       :implementation "QGIS:Simplify_Geometries"];
    :spatialOperation [a oso:Merge; ... ];
    :spatialOperation [a oso:asWKT ; ...];
  ] ;
  ...
  :linkingRule [:name "Link cyclone entities" ;
                :expression "prefix ....
                  construct {?emCyc owl:sameAs ?prevCyc} where ..." ;
  ] ;
  ... .
```

The :spatialOperation properties that appear in the rule description link every rule to the spatial operations it utilizes and that are defined in an ontology of spatial operations.

The ontology of spatial operations[10] contains geometry elements from GML, the Geometry Markup Language developed by the Open Geospatial Consortium[11] as well as spatial operations that apply on such geometries. As shown in figure 3, spatial operations are organised hierarchically and can have properties, such as *lossOfPrecision*.

The publication metadata can be used for several purposes, such as:

- precisely understanding how the published data have been generated
- assessing the accuracy of published data by examining the publication process, detecting operations that can potentially reduce the precision of the original source data;
- specifying new publications, adapted to new user needs, by modifying or augmenting the existing specification;

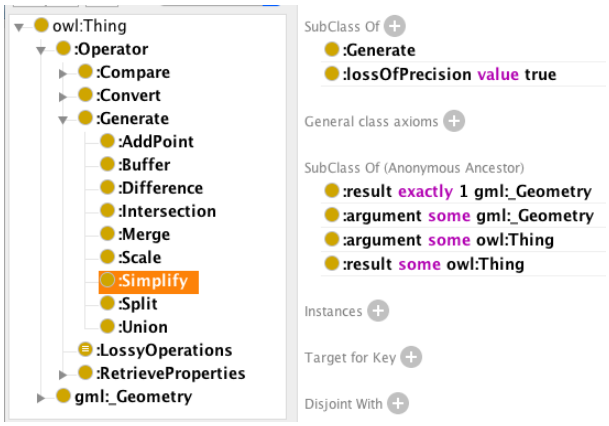
[9] These metadata have been created for this use case but are not (yet) published along with the data

[10] http://cui.unige.ch/isi/onto/GeometryAndOperations.owl

[11] http://www.opengeospatial.org/standards/gml

- testing and prototyping: it is easy to create small test RDF source datasets and to run SPARQL queries to test the production of published data. Once the mapping specification is correct, the implementation on the real source datasets can be implemented with these mappings as references.

Figure 3: Part of the ontology of spatial operations



5 Conclusion and Future Work

In this paper we showed that it is possible to formally specify the mapping and linking part of the linked geodata publication process by utilising SPARQL query expressions as a graph mapping rules. Such a specification provides an abstract description that is independent of the specific tools and systems that will be used to actually generate the published data. Therefore the data users can have a high level understanding of how the published data were obtained. They don't need to refer to technical tool or vendor specific documentation.

These specifications can easily be published as metadata that accompany the published linked data. Moreover, the mapping and linking rules are connected to an ontology of spatial operations. Therefore some properties of the published data (in terms of provenance, accuracy, or information loss) can be inferred by reasoning on the operations involved in the rules.

The experiment we carried out on a use case showed that this specification technique is sufficiently expressive to describe in a compact way a relatively complex mapping and linking process.

As mentioned in Section 4, the mapping and linking rules can be used to test and prototype the publication process. The next step, that we are currently studying, consists in directly using these specifications to produce the geodata publication scripts or applications. Of course, this is possible only when the functions involved in the specification do not rely on any human intervention. If human intervention is required then an interactive publication system must be generated.

References

Ehrig H., Ehrig K., Prange U., Taentzer G. (2006) *Fundamentals of Algebraic Graph Transformation*. Springer.

Garlik S. H., Seaborne A. (Eds.) (2013) SPARQL 1.1 Query Language. W3C Recommendation. [Online] Available from: https://www.w3.org/TR/sparql11-query/ [Accessed 14th February 2018].

Giuliani G. & Peduzzi P. (2011) The PREVIEW Global Risk Data Platform: a geoportal to serve and share global data on risk to natural hazards. *Natural Hazards and Earth System Science*, 11 (1):53-66.

Heath T. & Bizer C. (2011) Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1), 1-136.

Knublauch, H. (2013) SPIN - SPARQL Syntax. [Online] Available from http://spinrdf.org/sp.html [Accessed 2 April 2018],

Peduzzi P., Dao H. & Herold H. D. C. (2005) Mapping disastrous natural hazards using global datasets. *Natural Hazards*, 35(2):265-289.

Vilches-Blázquez LM., Villazón-Terrazas B., Corcho O., Gómez-Pérez A. (2014) Integrating geographical information in the Linked Digital Earth. *International Journal of Digital Earth*, 7 (7):554-575.

.