MidTerm Report

Publishing and Consuming Government Linked Data on the Semantic Web

Ghislain Auguste Atemezing

EURECOM-Multimedia Communications
Institut Mines-Télécom

December 19th, 2012

Supervisor:
Raphaël Troncy

EURECOM
Multimedia Department

Contents

Abstract

The need for geolocation is crucial for many applications for both human and software agents. More and more data is opened and interlinked using Linked Data principles, and it is worth modeling geographic data efficiently by reusing as much as possible from existing ontologies or vocabularies that describe both the geospatial features and their shapes. Our aim is to contribute to the actual efforts in representing geographic objects with attributes such as location, points of interest (POI), and addresses in the web of data, with a special focus on the French territory. As we publish data in RDF graphs, we are also aware of making them useful for the users. For that, we not only develop innovative applications to show up the value of data visualizations, but rather go beyond it. The challenge is to detect patterns to automatically develop an application using adequate visualization widgets in an affordable effort.

Chapter 1

Research Problems

1.1 Introduction

The Web is currently in a transition phase. After having been accessible on personal computers, it is now quickly moving to more and more ubiquity and entering in every part and moment of our lives. New devices and new ways to use them are being created. The ubiquity of the Web also creates an unseen abundance of information. Data is flowing onto the Web, created by users, generated by sensors, and stored in ever growing data farms.Geographic data is widely present on the web as they are used for location of Point of Interest. But those data lack of interoperability for a better integration due these three main factors:

- Vendor specific geometry support
- Different vocabularies , such as W3C Basic Geo, GML XMLLiteral, vendor-specific
- Different spatial reference systems, such as Lambert93, WGS84 , British National Grid, etc.

At the same time, many organizations are moving from legacy data stored in their databases to structured data on the web. Structured data is already present in the many databases, metadata attached to medias, and in the millions of spreadsheets created everyday across the world.

However, the recent emergence of linked data radically changes the way structured data is being considered. By giving standard formats for the publication and interconnection of structured data, linked data transforms the Web into a giant database. While making data available on the web, we need to build meaningful applications to show the value of all the huge data so that users could easily explore it, and derive new insights for it. As many information visualization tools are already present in InfoVis community[1], their easy adoption and usage for displaying structured data raise new challenges. Those challenges are two-folds:

- How to specify and define semantic web applications in terms of tools, widgets that can easily visualize RDF datasets?
- How to mine efficiently heterogeneous structured data to derive patterns for automatically recommend the adequate visualization tool to help users building innovative applications in an affordable time.

1.2 Contributions

The present report aims at giving a summary of our research we have been accomplishing during this first period of our thesis entitled: **"Publishing and Consuming Government Linked Data on the Semantic Web"**. Our concern was tackle the problematic in two directions :

- (i) - Geographic Information on the Web of data:as an application of the life-cycle of publishing geodata.
- (ii) - visualization tools for building innovative applications consuming structured data: as for leveraging the process of creating applications on-top of semantic data to highlight some relevant knowledge to the users.

1.2.1 Contributions on Geodata

Regarding this aspect of our research, we have achieved the following tasks :

- We have proposed an ontology describing features and point of interest for the French territory, by reusing existing taxonomy (GeOnto) and aligning it to other related vocabularies of the domain.
- We have made a comparative study of the triple stores, comparing their capability to store spatial information and their implementation of topological functions with respect to the ones existing in OGC[2] standards.
- We presented the output of those ideas in two conferences: the French Knowledge Engineering Community [1] and the Geographical Data at the Semantic Web Conference [2].

[1] http://en.wikipedia.org/wiki/Information_visualization

[2] http://www.opengeospatial.org/

1.2.2 Contributions on visualizations

Concerning our contributions on visualizations, we have contributed by :

- Building an application of the French first round elections using data from the data.gouv.fr and other public institutions.
- Building an application for conference events (confomaton) with their associated media reconcile from many social platform (instagram, twitter, etc.)
- Building a vocabulary for structuring applications on the Web of data
- Implementing a first version of a visualization module aiming at recommending a suitable tool for easing the creation of an application according to the dataset.

1.2.3 W3C Government Linked Data Working Group Contributions

We contribute to the W3C Government Linked Data Working Group (GLD WG)[3] activity since July, 2011. The objective of the Working Group is to *"provide standards and other information which help governments around the world publish their data as effective and usable Linked Data using Semantic Web technologies"*.

The group has three main task forces :

- **Task Force 1** aims to create a linked data community directory[4] and to maintain it on-line about deployments, vendors, contractors, end-user applications. In this work, we contribute to define the requirements and providing data for the French organizations in the directory.
- **Task Force 2** aims at providing **"Best Practices"** for Publishing Linked Data by producing recommendations regarding vocabulary selection, URI construction, Linked Data Cookbook, versioning, stability and provenance. Here, we are actively preparing a check list to help government to select and re-use vocabularies in their project. We have also proposed our vision of the Linked Open Data Lifecycle, best practices to construction URI and how to publish data which has multiple versions over a time period. We are also involved in defining relevant terms in the Linked Data Glossary[5] working Note. Apart from contributing in many sections of the document "Best Practices for Publishing Linked Data", we edited the working note [3].

[3] http://www.w3.org/2011/gld/

[4] http://dir.w3.org

[5] https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html

- **Task Force 3** goal is to provide relevant vocabularies to be used by governments or local authorities in their process of exposing their data. The scope of the vocabularies are: the description of people, business, data catalogue, organization, legal entity and statistical data.

Chapter 2

Modeling Geographic Information in LOD

2.1 Introduction

Many Linked Open Datasets have geospatial components, but still not having a common ways to describe features, spatial objects or geometries. Let us take the following two use-cases to express how difficult is to integrate geographic data of different datasets or sources.

```
UC1: What DBpedia Historic Buildings are within walking distance?
UC2: What OpenStreetMap Dog Parks are inside IGN
          Sophia-Antipolis Area?
```

Both use-cases take into account ”Concepts” (e.g: Historic Building, Dog Parks) that are defined differently depending on the provider of the dataset (e.g: DBPedia, OpenStreetMap, IGN). At the same time, the aforementioned Use-Cases implicitly make use of some specific topological functions widely use in the GIS applications, such as ”within” or ”inside”. Our aim is to contribute to the actual efforts in representing geographic objects to leverage the barrier of integration on the web of data. We focus on the French territory and we provide examples of representative vocabularies that can be used for describing geographic objects. We propose some alignments between various vocabularies (DBpedia, GeoNames, Schema.org, LinkedGeoData, Foursquare, etc.) in order to enable interoperability while interconnecting French geodata with other datasets. In France, there is currently a joint effort to publish geographic information in RDF and interlink them with relevant datasets. GeOnto is an ontology describing geospatial features for the French territory. We have proposed to align GeOnto with other popular vocabularies in the geospatial domain, using Silk[6] for schema mapping and we have evaluated the results. We studied how to extend the model to

[6] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

take into account efficient modeling for complex geometries. By doing so, tackle the complex geometry representation issues in the Web of Data, describing the state of implementations of geo-spatial functions in triple stores and comparing them to the new GeoSPARQL standard. We finally made some recommendations and advocate for the reuse of the NeoGeo ontology within GeOnto to better address the IGN requirements [2].

2.2 Geographic information in the Web of Data

2.2.1 LOD Cloud Review

The recent publication of statistics concerning the actual usage of vocabularies on the LOD cloud[7] provides not only an overview of best practice usage recommended by Tim Berners-Lee[8], but also provides a rapid view of the vocabularies re-used in various datasets and domains. Concerning the geographic domain, the results show that W3C Geo[9] is the most widely used vocabulary, followed by the `spatialrelations`[10] ontology of Ordnance Survey (OS). At the same time, the analysis reveals that the property `geo:geometry` is used in $1,322,302,221$ triples, exceeded only by the properties `rdf:type` ($6,251,467,091$ triples) and `rdfs:label`($1,586,115,316$ triples). This shows the importance of geodata on the web. Table tab:vocabLOD summarizes the results for four vocabularies (WGS84, OS spatial relation, Geonames ontology and OS admin geography) where the number of datasets using these vocabularies and the actual number of triples are computed.

Ontologies	#Datasets using	#Triples	SPARQL endpoint
W3C Geo	21	15 543 105	LOD cache
OS spatialrelations	10	9 412 167	OS dataset
Geonames ontology	5	8 272 905	LOD cache
UK administrative-geography	3	229 689	OS dataset

Table 2.1: Statistics on the usage of the four main geographic vocabularies (LOD cache should be understood as `http://lod.openlinksw.com/sparql/`). There are many more vocabularies used in the LOD cloud that contain also geographical information but that are never re-used.

[7] `http://stats.lod2.eu`

[8] `http://www.w3.org/DesignIssues/LinkedData.html`

[9] `http://www.w3.org/2003/01/geo/wgs84_pos`

[10] `http://data.ordnancesurvey.co.uk/ontology/spatialrelations`

2.2.2 Geodata Provider and Access

So far, the Web of data has taken advantage of geocoding technologies for publishing large amounts of data. For example, Geonames provides more than 10 millions records (e.g. 5, 240, 032 resources of the form `http://sws.geonames.org/10000/`) while LinkedGeoData has more than 60, 356, 364 triples. All the above mentioned data are diverse in their structure, the access point (SPARQL endpoint, web service or API), the entities they represent and the vocabularies used for describing them. Table 2.2 summarizes for different providers the number of geodata available (resources, triples) and how the data can be accessed.

Provider	#Geodata	Data access
DBpedia	727 232 triples	SPARQL endpoint
Geonames	5 240 032 (feature).	API
LinkedGeoData	60 356 364 triples	SPARQL endpoint, Snorql
Foursquare	n/a	API
Freebase	8,5MB	RDF Freebase Service
Ordnance Survey(Cities)	6 295 triples	Talis API
GeoLinkedData.es	101 018 triples	SPARQL endpoint
Google Places	n/a	Google API
GADM project data	682 605 triples	Web Service
NUTS project data	316 238 triples	Web Service
IGN experimental	629 716 triples	SPARQL endpoint

Table 2.2: Geodata by provider and their different access type.

2.3 Geodata Modeling Approach

2.3.1 Vocabularies for Features

Modeling of features can be grouped into four categories depending on the structure of the data, the intended purpose of the data modeling, and the (re)-use of other resources.

- (i) : One way for structuring the features is to define high level codes (generally using a small finite set of codes) corresponding to specific types. Further, sub-types are attached to those codes in the classification. This approach is used in the Geonames ontology[11] for codes and classes (A, H, L, P, R, S, T, U, V), with each of the letter corresponding to a precise category (e.g: A for administrative borders). Classes are then defined as `gn:featureClass a skos:ConceptScheme`, while codes are `gn:featureCode a skos:Concept`.

[11] `http://geonames.org/ontology/ontology_v3.0.rdf`

- (ii) : A second approach consists in defining a complete standalone ontology that does not reuse other vocabularies. A top level class is used under which a taxonomy is formed using the `rdfs:subClassOf` property. The LinkedGeoData ontology[12] follows this approach, where the 1294 classes are built around a nucleus of 16 high-level concepts which are: `Aerialway, Aeroway, Amenity, Barrier, Boundary, Highway, Historic, Landuse, Leisure, ManMade, Natural, Place, Power, Route, Tourism` and `Waterway`. The same approach is used for the French GeOnto ontology (Section 2.4), which defined two high-level classes `ArtificialTopographyEntity` and `NaturalTopographyEntity` with a total of 783 classes.

- (iii) : A third approach consists in defining several smaller ontologies, one for each sub-domain. An ontology network is built with a central ontology used to interconnect the different other ontologies. One obvious advantage of this approach is the modularity of the conceptualizing which should ease as much as possible the reuse of modular ontologies. Ordnance Survey (OS) follows this approach providing ontologies for administrative regions[13], for statistics decomposition[14] and for postal codes[15]. The `owl:imports` statements are used in the core ontology. Similarly, GeoLinkedData makes use of three different ontologies covering different domains.

- (iv) : A fourth approach consists in providing a *nearly flat list* of features or points of interest. This is the approach followed by popular Web APIs such as Foursquare types of venue[16] or Google Place categories[17]. For this last approach, we have built an associated OWL vocabulary composed of alignments with other vocabularies.

2.3.2 Vocabularies for Geometry Shape

The geometry of a point of interest is also modeled in different ways. We complete here the survey started by Salas and Harth [4] :

- *Point representation*: the classical way to represent a location by providing the latitude and longitude in a given coordinate reference system (the most used on the web is the WGS84 datum represented in RDF by the W3C Geo vocabulary). For example, Geonames defines the class `gn:Feature a skos:ConceptScheme` as a `SpatialThing` in the W3C Geo vocabulary.

[12] `http://linkedgeodata.org/ontology`

[13] `http://www.ordnancesurvey.co.uk/ontology/admingeo.owl`

[14] `http://statistics.data.gov.uk/def/administrative-geography`

[15] `http://www.ordnancesurvey.co.uk/ontology/postcode.owl`

[16] `http://aboutfoursquare.com/foursquare-categories/`

[17] `https://developers.google.com/maps/documentation/places/supported_types`

- *Rectangle* ("bounding box"): which represents a location with two points or four segments making a geo-referenced rectangle. In this way of modeling, the vocabulary provides more properties for each segment. The FAO Geopolitical ontology[18] uses this approach.

- *List of Points*: the geometry shape is a region represented by a collection of points, each of them being described by a unique RDF node identified by a lat/lon value. The `Node` class is used to connect one point of interest with its geometry representation. The POI are modeled either as `Node` or as `Waynode` (surfaces). This approach is followed by LinkedGeoData [5]. Another way is to represent the geometry shape by a group of RDF resources called a "curve" (similar to LineString of GML). This approach is the one used in GeoLinkedData [6].

- *Literals*: the vocabulary uses a predicate to include the GML representation of the geometry object, which is embedded in RDF as a literal. This approach is followed by Ordnance Survey [7].

- *Structured representation*: the geometry shape is represented as a typed resource. In particular, polygons and lines are represented with an RDF collection of basic W3C Geo points. This approach is used by the NeoGeo vocabulary[19].

In [1], we provide a scenario from DBpedia, Geonames and LinkedGeoData to give an overview of the different models used by the provider to depict the same "reality", which is the district of the 7th Arrondissement in Paris. Regarding the "symbolic representation", two datasets opted for "Feature" (DBpedia and Geonames) while LGD classifies it as a "Suburb" or "Place". They all represent the shape of the district as a POINT which is not very efficient if we consider a query such as *show all monuments located within the 7th arrondissement of international importance.* To address this type of query and more complicated ones, there is a need for more advanced modeling as we describe in the next section.

2.4 Aligning Geo Vocabularies

The purpose of alignment is to reconcile differences in data semantics and facilitate the linking process across existing spatial data resources. The scope of the source vocabulary is the French territory with IGN, the public institution for providing geographic data. IGN is also experimenting in exposing some of their data as Linked Data and acts as an important provider in the `http://data.gouv.fr` portal.

[18] `http://www.fao.org/countryprofiles/geoinfo/geopolitical/resource/`

[19] `http://geovocab.org/doc/neogeo/`

2.4.1 Existing Vocabularies

IGN has developed two complementary vocabularies (GeOnto and bdtopo) which differ in their provenance but have the same scope, which is to describe geographic entities in the French territory. GeOnto is the product of a research project[20] aiming at building and aligning heterogeneous ontologies in the geographic domain. The "light" version of the final ontology[21] defines two top classes for a total of 783 classes and 17 properties (12 DP / 5 OP). GeOnto has labels in both French and English, but has no comments specified for the resources. The bdtopo ontology is derived from a geospatial database with the same name. It contains 237 classes and 51 properties (47 DP / 4 OP). All the labels and comments are in French.

2.4.2 GeOnto Alignment Process

We choose GeOnto because it covers a large number of categories and also has labels in English. We have performed the alignment with five OWL vocabularies (bdtopo, LGD, DBpedia, Schema.org and Geonames) and two flat taxonomies (Foursquare, Google Place). For the latter, we have transformed the flat list of types and categories into an OWL ontology. For each alignment performed, we only consider `owl:equivalentClass` axioms. We use the Silk tool [8] to compute the alignment using two metrics for string comparison: the *levenshteinDistance* and *jaro* distances. They work on the English labels except for the alignment with bdtopo where we use the French labels. We apply the average aggregation function on these metrics with an empirically derived threshold. However, for generating the final mapping file for vocabularies of small size, we manually validate and insert relations of type `rdfs:subClassOf`. The threshold to validate the results is set to 100% for links considered to be correct and greater than 40% for links to be verified. The alignment with Geonames is special, considering the property restriction used in the ontology for codes.

Table 2.3 summarizes the result of the alignment process between GeOnto and the existing vocabularies/taxonomies. All the resources of this work are available at `http://semantics.eurecom.fr/datalift/tc2012/`.

In general, we obtain good results with Silk, with precision beyond 80%: Google Place: 94%, LGD: 98%, DBpedia: 89%, Foursquare: 92% , Geonames: 87% and bdtopo: 92%. We obtained a precision of only 50% with schema.org due to numerous fine-grained categories that are badly aligned (e.g. `ign:Berge owl:equivalentClass schema:Park`).

[20] `http://geonto.lri.fr/Livrables.html`

[21] `http://semantics.eurecom.fr/datalift/tc2012/vocabs/GeoOnto/`

Vocabulary	**#Classes**	**#Aligned Classes**
LGD	`owl:Class`:1294	178
DBpedia	`owl:Class`:366	42
Schema.org	`owl:Class`:296	52
Geonames	`skos:ConceptScheme`:12	–
	`skos:Concept`:699	287
Foursquare	359	46
Google Place	126	41
bdtopo	`owl:Class`:237	153

Table 2.3: Results of the alignment process between GeOnto and existing vocabularies/taxonomies.

2.5 GeoSPARQL Standard

OGC has adopted the GeoSPARQL standard to support both representing and querying geospatial data on the Semantic Web. The standard document [9] contains 30 requirements. It also defines a vocabulary for representing geospatial data in RDF and provides an extension to the SPARQL query language for processing geospatial data. The proposed standard follows a modular design with five components: (i) A *core component* defining top-level RDFS/OWL classes for spatial objects; (ii) a *geometry component* defining RDFS data types for serializing geometry data, RDFS/OWL classes for geometry object types, geometry-related RDF properties, and non-topological spatial query functions for geometry objects; (iii) a *geometry topology component* defining topological query functions; (iv) a *topological vocabulary component* defining RDF properties for asserting topological relations between spatial objects; and (iv) a *query rewrite component* defining rules for transforming a simple triple pattern that tests a topological relation between two features into an equivalent query involving concrete geometries and topological query functions. Each of the components described above has associated requirements. Concerning the vocabulary requirements, the document specifies seventeen requirements presented in the GeoSPARQL draft document.

Based on the GeoSPARQL requirements, we were interested in comparing some geospatial vocabularies[22] to see how far they take already into account topological functions and which are the standard they followed among OpenGIS Simple Features (SF), Region Connection Calculus (RCC) and Egenhofer relations. We find that the NeoGeo (Spatial and Geometry) and OS Spatial vocabularies have integrated in their modeling partial or full aspects of topological functions as summarized.

As geodata has to be stored in triple stores with efficient geospatial in-

[22] `http://lov.okfn.org/dataset/lov/details/vocabularySpace_Geography.html`

dexing and querying capabilities, we also survey the current state of the art in supporting simple or complex geometries and topological functions compatible with SPARQL 1.1. Table 2.4 shows which triple stores can support part of the GeoSPARQL standard regarding serialization and spatial functions.

2.6 Discussion

The alignment of GeOnto provided in the previous section enables interoperability of symbolic descriptions. The need for a better choice of geometric structure, typically the choice between literal versus structured representations depends on three criteria: (i) the coverage of all the complex geometries as they appear in the data; (ii) a rapid mechanism for connecting "features" to their respective "geometry"; (iii) the possibility to serialize geodata into traditional formats used in GIS applications (GML, KML, etc.) and (iv) the choice of triple stores supporting as many as possible functions to perform quantitative reasoning on geodata. It is clear that a trade-off should be taken depending on the technological infrastructure (e.g: data storage capacity, further reasoning on specific points on a complex geometry).

- **Complex Geometry Coverage:** We have seen that on the Web of Data, there are few modeling of geodata with their correct shape represented as a LINE or POLYGON. However, some content providers (e.g. IGN) need to publish all types of geodata including complex geometries representing roads, rivers, administrative regions, etc. Two representations are suitable: *OS Spatial* and *NeoGeo* ontologies . Direct representation of the GeoSPARQL vocabulary is also suitable.

- **Features connected to Geometry:** In modeling geodata, we advocate a clear separation between the features and their geometry. This is consistent with the consensus obtained from the different GeoVocamps[23] and the outcome of this approach is expressed in the modeling design of NeoGeo. The top level classes `spatial:Feature` and `geom:Geometry` are connected with the property `geom:geometry`.

- **Serialization and Triple stores:** We also advocate the use of properties that can provide compatibility with other formats (GML, KML, etc.). This choice can be triple store independent, as there could be ways to use content-negotiation to reach the same result. In Table 2.4, `Open Sahara`[24], `Parliament` [25], `Virtuoso`[26] are WKT/GML-compliant with respectively 23 and 13 functions dealing with geodata.

[23] http://www.vocamp.org

[24] http://www.opensahara.com

[25] http://geosparql.bbn.com

[26] http://www.openlinksw.com

Triple store	WKT	GML	Geometry	Functions	GeoVocab
Virtuoso	Yes	Yes	Point	13	W3C Geo Typed Literal
Allegro-Graph	-	–	Point	3	“strip” mapping data
OWLIM-SE	–	–	Point	4	W3C Geo
Open Sahara	Yes	Yes	Point, Line, Polygons	23	Typed Literal
Parliament	Yes	Yes	Point, Line, Polygons	23	GeoSPARQL

Table 2.4: Triple stores survey with respect to geometry types supported and geospatial functions implemented.

- **Literal versus structured Geometry:** Decomposing a LINE or a POLYGON into multiple results in an “explosion” in the size of the dataset and the creation of numerous blank nodes. However, sharing points between descriptions is a use case with such a need. IGN has such use-cases and the natural solution at this stage is to consider reusing the NeoGeo ontology in the extended version of GeOnto. The choice of the triple store (e.g.,Virtuoso vs Open Sahara) is not really an issue, as the IndexingSail[27] service could also be wrapped on-top of Virtuoso to support full OpenGIS Simple Features functions[28].

[27] https://dev.opensahara.com/projects/useekm/wiki/IndexingSail

[28] http://www.opengeospatial.org/standards/sfs

Chapter 3

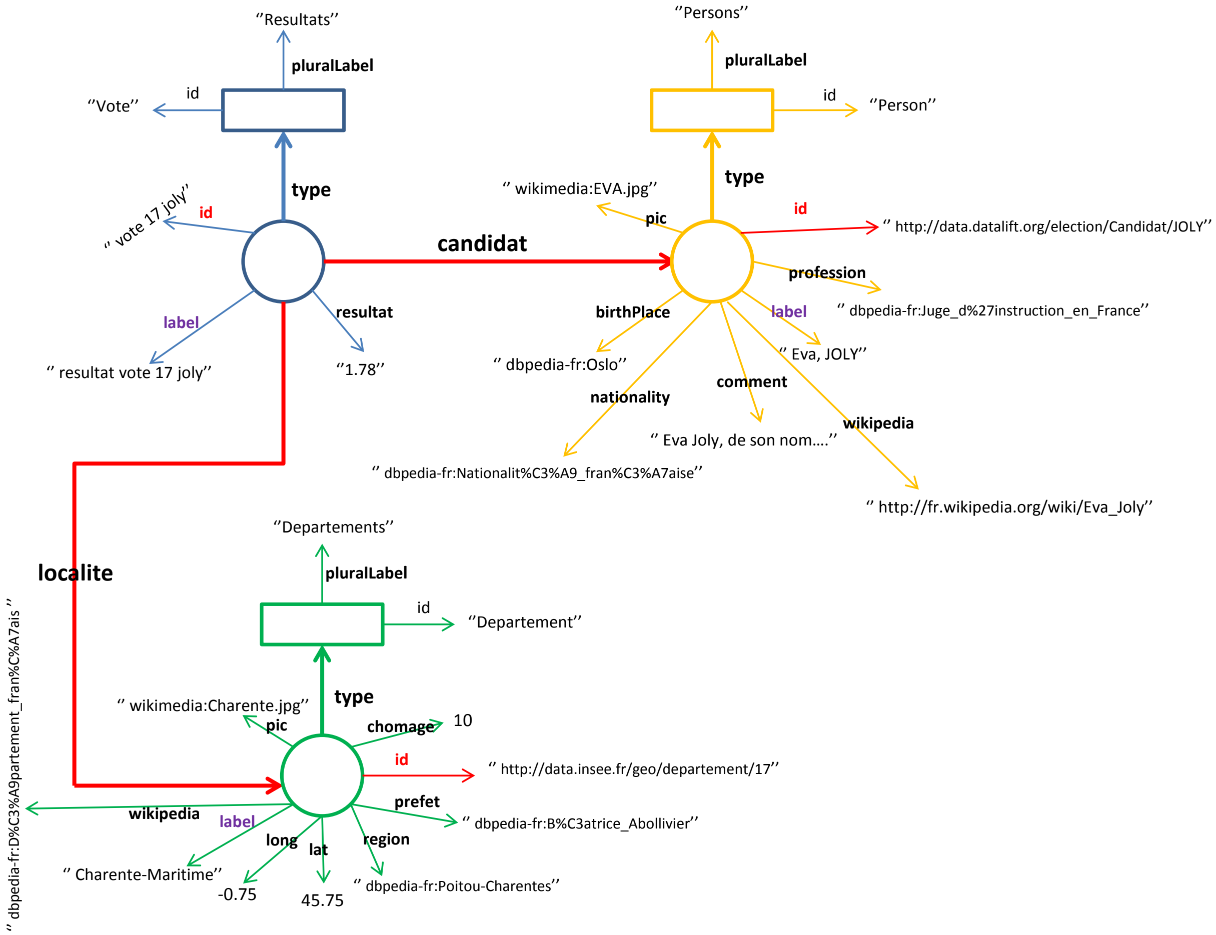
Visualization Tools in Linked Government Data

3.1 Introduction

We first review the numerous applications that have been developed on top of datasets that have been opened by governments (UK, USA, France) and local authorities. We have then derived and proposed height use cases that can be developed to consume data from the different main data providers in France: INSEE, DILA, IGN, FING, etc. We mention that the most interesting Use Cases (UCs) are the ones which show the added value of having interconnected datasets. These UCs, developed and deployed, can be useful to show the benefits of Linked Data in a variety of domains such as Education, Tourism, Cultural Heritage, Civil administrations, Judicial Court, Medicine, etc. As a good starting point, we have developed an application[29] for the first round of the French election reusing five different datasets, as Figure 3.1 shows the data model used for the design of the application. Moreover, we observed that many successful applications that have been developed visualize structured data around the geographical, time and concepts dimensions. Furthermore, we analyze some requirements expected from the users and actors/providers of Open Data in France. The abovementioned use-cases are currently tested and validated within the Datalift[30] Platform [10].

[29] http://www.eurecom.fr/~atemezin/DemoElection/

[30] http://www.datalift.fr



"Resultats"
pluralLabel
id
"Vote"
type
id
" vote 17 joly"
candidat
label
resultat
" resultat vote 17 joly"
"1.78"
"Persons"
pluralLabel
id
"Person"
type
" wikimedia:EVA.jpg"
pic
id
" http://data.datalift.org/election/Candidat/JOLY"
profession
" dbpedia-fr:Juge_d%27instruction_en_France"
birthPlace
label
" dbpedia-fr:Oslo"
" Eva, JOLY"
nationality
comment
wikipedia
" Eva Joly, de son nom…."
" dbpedia-fr:Nationalit%C3%A9_fran%C3%A7aise"
" http://fr.wikipedia.org/wiki/Eva_Joly"
localite
"Departements"
pluralLabel
id
"Departement"
type
" wikimedia:Charente.jpg"
pic
chomage
10
id
" http://data.insee.fr/geo/departement/17"
wikipedia
label
prefet
" dbpedia-fr:B%C3atrice_Abollivier"
long
lat
region
" Charente-Maritime"
-0.75
45.75
" dbpedia-fr:Poitou-Charentes"
" dbpedia-fr:D%C3%A9partement_fran%C%A7ais"

Figure 3.1: Exhibit Data model of the score obtained by candidate Eva Joly in Charente-Maritime, linked with knowledge from DBpedia, INSEE and Wikipedia.

Regarding tools used for visualization, we have divided them into two categories, providing for each of them relevant examples: (i)-tools that operate over RDF data, and (ii) tools that operate over other structured format. We then provide some basic criteria for assessing a given visualization tool, with some weight attached to each of the criterion.

3.2 Usage Scenarii for Application

The most interesting use cases for Government Open Data are the ones which show the added value of having interconnected datasets. The need for interlinking data is almost never spontaneously expressed by the actors in France. During our interviews, only a few of them did express it, namely CG 33[31], Etalab[32] and Montpellier[33]. However, what is expressed, is the desire for interoperability and ease of cross-linking. Interoperability allows developers to not duplicate efforts, thus for example not duplicating a connector each time one is changing his geographical region. We list in [11] ten key functionalities expected from the users and actors/providers of Open Data in France. It is clear that there exits some business needs that can be summarized as:

- `Access by territory`: case of Direction of Legal Information in France[34]
- `Access by time period`: History, genealogy ,anecdotal history (e.g: *It happens someday in..*), temporal search, temporal inquiry, games, etc.

3.3 Tools for Visualizing Data on the Semantic Web

We have studied a number of tools (14) used for visualizing raw data or RDF data. The main objective is to select which could be appropriate for an agile prototyping and development of an application consuming heterogeneous data. The study in [12] describes each tool according to the data formats accepted as input, how the data is accessed (API or SPARQL endpoint), the programming language of the tool, the type of views available (e.g: chart, map, text, tabular, et.), the list of libraries embedded, the license and the creator (whether organization, project or a single person). The outcome of this state-of-the-art can then be used to assess in the choice of a given visualization tool, according to some criteria, such as (i) usability, (ii) visualization, (iii) data accessibility, (iv) deployment and (v) extensibility.

[31] `http://datalocale.fr`

[32] `http://www.etalab.gouv.fr`

[33] `http://opendata.montpelliernumerique.fr`

[34] `http://www.dila.premier-ministre.gouv.fr/`

3.4 DVIA: A vocabulary Describing VIsualization Application

DVIA is a small vocabulary aims at describing any applications developed to consume datasets in 4-5 stars, using visual tools to showcase the benefits of Linked Data. It reuses four existing vocabularies: `Dublin Core terms` [35], `dataset catalogue (DCAT)`[36], `Dublin Core Metadata Initiative`[37] and `Organization vocabulary`[38]. It is composed of three main classes :

- **Application**: This class represents the application or the mashup developed for demo-ing or consuming data in LD fashion. It is subclass of **dctype:Software**
- **Platform**: The platform where to host or use the application, could be on the web (Firefox, chrome, IE, etc..) or mobile (android, etc..) or event desktop
- **VisualTool**: Represents the tool or library used to build the application.

The diagram of the main classes and properties is depicted in Figure 3.2. The goal is to test the vocabulary with all the finalists applications submitted at the Semantic Web Challenge in the two previous years (2011-2012). The actual version of the vocabulary in turtle format can be found here http://www.eurecom.fr/ atemezin/datalift/visumodel/visu-vocab.ttl directory, along with a sample description of the application which won the Semantic Web Challenge[39] this year[40].

[35] `http://purl.org/dc/terms/`

[36] `http://www.w3.org/ns/dcat#`

[37] `http://purl.org/dc/dcmitype`

[38] `http://www.w3.org/ns/org#`

[39] urlhttp://challenge.semanticweb.org/2012/winners.html

[40] `http://www.eurecom.fr/~atemezin/datalift/visumodel/eventMedia-sample.ttl`

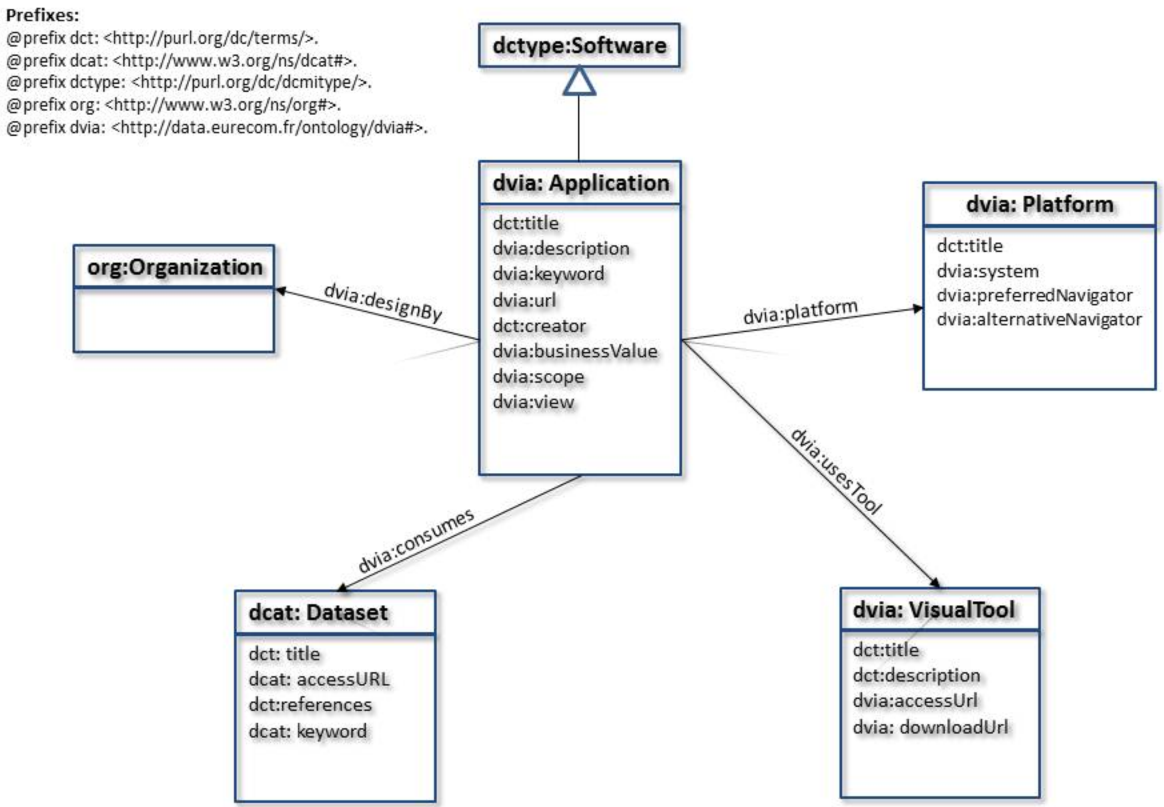


Figure 3.2: Conceptual Model of the DVIA vocabulary

3.5 Examples of Visualization

We had the opportunity to use some of the tools described in [12] to build some innovative applications using heterogeneous data exposed as RDF. We describe two of them in the following subsections.

3.5.1 French Election Application

The application was built to showcase the use of a visualization application of top of heterogeneous government data exposed in RDF. Election[41] is a mashup implemented in Exhibit using four datasets: the first results of presidential election in France, the rate of unemployment by INSEE, fr.dbpedia (http://fr.dbpedia.org/), dbpedia.org and geolocation data of districts and regions by IGN. The application provides means to search by a picture of a candidate, and make use of filters representing five dimension facets (candidate, region, department, rate of vote, rate of unemployment) to depict on a map centered to France, the appropriate results, along with more relevant details about the candidate, the description of the region, etc... Figure 3.1 shows a sample of the data model follows to implement the back-end of the application.

3.5.2 Confomaton Application

Confomaton [13] is a semantic web application that aggregates and reconciles information such as tweets, slides, photos and videos shared on social media that could potentially be attached to a scientific conference. The main demonstrator is available at *http://eventmedia.eurecom.fr/confomaton* reflecting the up-to-date conferences coming from Lanyard feeds. A second demonstrator corresponding to the archived of ISWC 2011 conference [14] is available at http://eventmedia.eurecom.fr/iswc2011. The user interface for the aforementioned demonstrators is built around four perspectives (tabs in the UI) characterizing an event: *(i) Where does the event take place?*, *(ii) What is the event about?*, *(iii) When does the event take place?*, and finally *(iv) Who are the participants of the event?*. Those four perspectives correspond to views in the Confomaton API configuration.

[41] `http://www.eurecom.fr/~atemezin/DemoElection/`

Chapter 4

Conclusions and Future work

We have presented in this document our main contributions in some issues around publishing and consuming Government Linked Data on the Semantic Web. We have first focus on geographic data, proposing some best practices to take into account for better interoperability of features and geometry. We apply our propositions on the French domain, starting from existing vocabulary to mapping with other similar vocabularies. The goal is to leverage the alignment of the dataset generated with other providers in the LOD cloud. In the second part of the document, we have studied scenarii and tools be used for creating visual applications to better get insights of the data. We propose a small vocabulary to describe visualization applications as a first step to formally describe as much as possible applications developed in top of government datasets, to improve their re-usability. However, we have many challenges to tackle in the following months to achieve all the goals of our problematic.

Regarding the geodata modeling, our future work includes the conversion and publication of a large RDF dataset of geographic information of the French territory together with alignments with other datasets at the instance level. At the same time, we plan to publish with IGN a new version of an adequate ontology for describing features and geometry according some best practices we are contributing to elaborate. Furthermore, some alignments with the new OGC standard GeoSPARQL will be performed and evaluated.

Regarding the visualization tools, some further studies should be made for mobile applications, as they are not considered in this current study. We plan also to develop a small vocabulary that could be used to describe OpenData applications. At the same time, we will develop more applications using different datasets for detecting patterns for visualizing RDF data. For this challenge, we will need to study the underlying data (list of properties, number of triples, categories, etc.), the ontologies used, the templates or libraries for visualizations (Exhibit, GeoAPI, LDA, Sparkl, d3.js,etc) and finally the effort for a user to build the application. This could lead to a framework for generating automatically visualizations from heterogeneous datasets extracting relevant features by data mining techniques.

We also plan to expose as RDF some relevant applications submitted at the Semantic Web Challenge the past ten years using the DVIA vocabulary. The idea is to have a hub of a dataset describing the tools, datasets, etc. used to build applications consuming 4-5 stars datasets.

Bibliography

[1] G.A. Atemezing and R. Troncy. Vers une meilleure interopérabilité des données géographiques françaises sur le web de données. *Actes des 23es Journées Francophones d'Ingénierie des Connaissances (IC 2012)*, pages 369–384, 2012.

[2] G.A. Atemezing and R. Troncy. Comparing vocabularies for representing geographical features and their geometry. *Terra Cognita 2012 Workshop*, page 3, 2012.

[3] D. Wood, G. Atemezing, and B. Hyland. Common glossary of terms in linked data. In *W3C Government Linked Data Working Group Note*. 2012.

[4] Juan Salas and Andreas Harth. Finding spatial equivalences accross multiple RDF datasets. In *Terra Cognita Workshop*, pages 114–126, Bonn, Germany, 2011.

[5] Sören Auer, Jens Lehmann, and Sebastian Hellmann. LinkedGeoData - Adding a Spatial Dimension to the Web of Data. In *International Semantic Web Conference (ISWC'09)*, 2009.

[6] Alexander de León, Luis M. Vilches, Boris Villazón-Terrazas, Freddy Priyatna, and Oscar Corcho. Geographical linked data: a Spanish use case. In *International Conference on Semantic Systems (I-SEMANTICS'10)*, Graz, Austria, 2010.

[7] John Goodwin, Catherine Dolbear, and Glen Hart. Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web. *Transactions in GIS*, 12:19–30, 2008.

[8] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and Maintaining Links on the Web of Data. In *International Semantic Web Conference (ISWC'09)*, 2009.

[9] Matthew Perry and John Herring. OGC GeoSPARQL- A Geographic Query Language for RDF Data. In *OGC Implementation Standard, ref: OGC 11-052r4*, 06 2012.

[10] F. Scharffe, G. Atemezing, R. Troncy, F. Gandon, S. Villata, B. Bucher, F. Hamdi, L. Bihanic, G. Képéklian, F. Cotton, J. Euzenat, Z. Fan, PY. ; Vandenbussche, and B. Vatant. Enabling linked-data publication with the datalift platform. *In AAAI 2012, 26th Conference on Artificial Intelligence, W10:Semantic Cities, July 22-26*, 2012.

[11] G. Atemezing, C. Nepote, and R. Troncy. Usage scenarii for applications (v.1.1). In *Deliverables 6.1 of DataLift*. 2012.

[12] G. Atemezing and R. Troncy. Tools for visualization (v.1.2). In *Deliverables 6.2 of DataLift*. 2012.

[13] H. Khrouf, G. Atemezing, T. Steiner, G. Rizzo, and R. Troncy. Confomaton: A conference enhancer with social media from the cloud. *In 9th Extended Semantic Web Conference (ESWC'12), Demo Session*, 2012.

[14] H. Khrouf, G. Atemezing, T. Steiner, G. Rizzo, and R. Troncy. Aggregating social media for enhancing conference experience. *In (ICWSM'12) 1st International Workshop on Real-Time Analysis and Mining of Social Streams (RAMSS'12)*, pages 34–37, 2012.

[15] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition, 2011.

[16] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5:1–22, 2009.

[17] Sébastien Mustière, Nathalie Abadie, Nathalie Aussenac-Gilles, Marie-Noelle Bessagnet, Mouna Kamel, Eric Kergosien, Chantal Reynaud, and Brigitte Safar. GéOnto : Enrichissement d'une taxonomie de concepts topographiques. In *Spatial Analysis and GEOmatics (Sageo'09)*, Paris, France, 2009.

[18] Krzysztof Janowicz, Sven Schade, Arne Bröring, Carsten Kessler, Christoph Stasch, Patrick Maué, and Thorsten Diekhof. A transparent semantic enablement layer for the geospatial web. In *Terra Cognita Workshop*, 2009.

[19] Jean-Daniel Fekete, Jarke J. Wijk, John T. Stasko, and Chris North. The value of information visualization. *In Information Visualization: Human-Centered Issues and Perspectives*, pages 1–18, 2008.

[20] Ed H. Chi. A taxonomy of visualization techniques using the data state reference model. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, INFOVIS '00, pages 69–, Washington, DC, USA, 2000. IEEE Computer Society.