Towards a Linked Dataset for French Districts Evolution

Ghislain Auguste Atemezing
MONDECA
35 boulevard Strasbourg, Paris, France
ghislain.atemezing@mondeca.com

ABSTRACT

This paper provides a real-world scenario of combining heterogeneous datasets in plain text and shape files to create a dataset in RDF of the French districts evolution since 1790. We use two different datasets (1) a shape file containing polygons from the French Geographic Institute (IGN) and (2) a text file containing names of districts and events occurred at a certain time (fusion, merging, etc) with other entities from the National Statistics Institute (INSEE).

This work explores the use of semantic technologies to combine heterogeneous datasets (text and shape) for creating an RDF dataset explaining the geo-dynamic evolution of districts occurred since the French revolution. The resulting dataset reuses standard vocabularies for topographic entities, geometries, provenance and time. The expected outcome of the dataset is to interlink with other historical facts and build innovative applications consuming the dataset.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.5 [**Online Information Services**]: Data sharing—*Web-based services*

General Terms

geospatial data, RDF modeling, Linked Data

Keywords

geodata, Linked dataset, aligning heterogeneous data

1. INTRODUCTION

So far, Linked Data principles and practices are being adopted by an increasing number of data providers, getting as result a global data space on the Web containing hundreds of LOD datasets [4]. There are already several guidelines for generating, publishing, interlinking, and consuming Linked Data [4]. An important task, within the generation process, is to build the vocabulary to be used for modelling the domain of the data sources, and the common recommendation is to reuse as much as possible available vocabularies [4, 5]. This reuse approach speeds up the vocabulary development, and therefore, publishers will save time, efforts, and resources.

There are research efforts, like the NeOn Methodology [9], the Best Practices for Publishing Linked Data - W3C Working Group Note [5], and the work proposed by Lonsdale et al. [8]. However, at the time of writing we have not found specific and detailed guidelines that describe how to reuse available vocabularies at fine granularity level,i.e., reusing specific classes and properties. Our claim is that this difficulty in how to reuse vocabularies at fine grained level is one of major barriers to the reuse of vocabularies on the Web and in consequence to deployment of Linked Data.

Moreover, the recent success of Linked Open Vocabularies (LOV[1]) as a central point for curated catalog of ontologies is helping to convey on best practices to publish vocabularies on the Web, as well as to help in the Data publication activity on the Web. LOV comes with many features, such as an API, a search function and a SPARQL endpoint.

In this paper we propose a set of guidelines for this task, and provide technological support by means of a plugin for Protégé , which is one of the popular frameworks for developing ontologies in a variety of formats including OWL, and RDF(S). It is backed by a strong community of developers and users in many domains. One success on Protégé also depends on the availability to extend the core framework adding new functionalities by means of plug-ins. In addition, we propose to explore, design and implement a plug-in of LOV in Protégé for easing the development of ontologies by reusing existing vocabularies at fine grained level. The tool helps to improve the modeling and reuse of ontologies used in the LOD cloud.

2. RELATED WORK

{TODO: I suggest to move this related work at the end**}**
In the literature there exist many attempts to advise vocabulary publishers on the importance of reusing terms, as indicated in [6, 7]. However, to the best of our knowledge there are not guidelines to help vocabulary practitioners to reuse vocabularies in real-world scenario, and considering specific ontology/vocabulary elements.

In the W3C Government Linked Data best practice document [5], reusing vocabularies is recommended by providing to stakeholders a basic checklist when using or extending a

[1] `http://lov.okfn.org/dataset/lov/`

vocabulary. It gives general guidance to follow before publishing the vocabulary, not guidance during the creation of the vocabulary. Our proposal is to guide the users during the implementation process.

Recently, Janowicz et al. have proposed a 5 stars rating for Linked (Open) Data vocabulary use to "encourage data owners, engineers and practitioners to publish and use vocabularies on the Web" [6]. They make it clear that the rating do not refer to the quality of the vocabularies. In the definition of the rating system, the third star is given to a vocabulary linked to other vocabularies by means of explicit alignments and import of external vocabularies. Our guidelines make it easier to vocabulary publishers to obtain at least 3-star vocabularies.

Other initiatives similar to the tool we have developed can be found in the literature but not currently maintained. The BioPortal Reference Plugin[2] allows the user to insert into the ontology references to external ontologies and terminologies stored in BioPortal[3]. The plugin allows to generate external reference of a selected term. Additionally, the BioPortal Import Plugin[4] allows users to import classes from external ontologies stored in the BioPortal ontology repository. The user can import entire trees of classes with a desired depth and choose which properties to import for each class. However, those plugins work only with Protégé 3.x releases and are not ported yet to recent versions.

Most closely related to the Protégé LOV plugin are approaches that use semantic search engine to support the process of editing an ontology and make large scale knowledge reuse automatically integrated in the tool. An example is the Watson Plugin [2] for the NeOn Toolkit [3], a plugin supporting the NeOn life-cycle management using the Watson [1] APIs[5].However the similar plugin for Protege[6] is just a proof of concept rather than a real plugin.

3. LEGACY DATASETS

We obtained the dataset from two different sources: the text file from INSEE website in text file and the shape file from the output of an ANR project.

3.1 Text file with districts events

The text from INSEE contains information of different modifications occurred in France from 1930 until 2013. Each line represents a piece of information with an INSEE code (5 digit), the date in the form DD/MM/YYYY and a string describing the event. For example, " 38205 09/05/1947 Lans devient Lans-en-Vercors" states that on the 9th May, 1947 Lans with INSEE code 38205 became "Lans-en-Vercors".

3.2 Shape file with geometries

We obtained a license copy of the shape file with the ANR project Geopeuple[7]

In a nutshell, the activity of building vocabularies by reusing available vocabulary terms consists of the following tasks:

Figure 1: Workflow for reusing available terms when building ontologies

- Search for suitable vocabulary terms to reuse from any vocabulary repository, such as BioPortal[8], LOV[9], Bartoc.org[10], etc. The search should be conducted by using the terms of the application domain.
- Assess the set of candidate terms from the vocabulary repository. The goal of this task is to find out if the candidate terms are useful and relevant for the ontology being built. For this purpose we can rely on metrics provided by the vocabulary repository. In the particular case of LOV, the results include a score related to their "importance" in the corpus for each term retrieved. For every vocabulary in the LOV, terms (classes, properties, datatypes, instances) are indexed and a full text search feature is offered[11]. The Linked Open Vocabularies search engine ranking algorithm takes into account the popularity of the term within the LOV ecosystem and most importantly assigned a different score depending on which label property a searched term matched [10].
- Select the most appropriate term taking into account the assessment task. To distinguish between those candidate terms which are the most suitable, we propose to use the following criteria (i) the stability of the URI namespace, (ii) the trustworthiness of the publisher of the vocabulary and (iii) the presence or absence of a community using the vocabulary. With respect to LOV repository we can also rely on the score of terms.
- Include the selected term in the ontology that has being developed. There are at least three alternatives in this case
  - Include the selected term and use it directly, i.e., as it is, in the local ontology by defining local axioms to or from that term in the local ontology.
  - Include the selected term, create a local term, and define the `rdfs:subClassOf` or `rdfs:subPropertyOf` axiom to relate both terms.
  - Include the selected term, create a local term, and define the `owl:equivalentClass` or `owl:equivalentProperty` axiom to relate both terms.

It is worth mentioning the following considerations

- We want to promote a "responsible" reutilization, i.e., if the user already found something interesting in a particular vocabulary, the user would probably like to dig a little bit more on it, suggesting other terms in it, before jumping somewhere else.
- We want to keep consistency of reused terms . Therefore, it would be necessary to (1) check and update the domain/range of the selected terms; (2) change cardinalities; and (3) add some restrictions.

[2] `http://protegewiki.stanford.edu/wiki/BioPortal_Reference_Plugin`
[3] `http://bioportal.bioontology.org/`
[4] `http://protegewiki.stanford.edu/wiki/BioPortal_Import_Plugin`
[5] `http://watson.kmi.open.ac.uk/WS_and_API.html`
[6] `http://protegewiki.stanford.edu/wiki/Watson_Search_Preview`
[7] `http://geopeuple.ign.fr/`
[8] `http://bioportal.bioontology.org/`
[9] `http://lov.okfn.org/`
[10] `http://bartoc.org`
[11] `http://lov.okfn.org/dataset/lov/terms`

- We want to bring as many knowledge as possible from the reused term, e.g., for a particular property it would be necessary to check if a inverse property exists and import it as well.

4. RDF DATASET CREATION

Include the process of creating RDF dataset

The intended purpose of the LOV [10] is to help users to find and reuse terms of vocabularies in Linked Open Data. For achieving that purpose, the LOV gives access to vocabularies metadata and terms using programmatic access with APIs. LOV[12] catalogue is a hub of curated vocabularies used in the Linked Open Data Cloud, as well as other vocabularies suggested by users for their reuse. Some of the three main features of the LOV are for: (1) searching ontologies according to their scope, (2) assessing ontologies by providing a score for each term retrieved by a keyword search and (3) interconnecting ontologies using VOAF vocabulary[13]. The search function uses an algorithm based on the term popularity in LOD and in the LOV ecosystem. The matched resource set for each term should be the value of the a) `rdfs:label` b) `rdfs:comment`, or c) `rdfs:description` property.

For every vocabulary in the LOV, terms (classes, properties, datatypes, instances) are indexed and a full text search feature is offered[14]. The Linked Open Vocabularies search engine ranking algorithm takes into account the popularity of the term within the LOV ecosystem and most importantly assigned a different score depending on which label property a searched term matched. For example, a term matching a value for the property `rdfs:label` will have a higher score than if it matches a value for the property `dcterms:comment`.

Futhermore, the LOV APIs give a remote access to the many functions of LOV through a set of RESTful services[15]. The APIs give access through three different type of services related to: (1) vocabulary terms (classes, properties, datatypes and instances), (2) vocabulary browsing and (3) ontology's creators.

Acknowledgments..

Thanks to Pierre-Yves and the LOV team for maintaining the LOV catalog and the API access. This work has been supported by the KDrive Project.

5. REFERENCES

[1] M. D'Aquin, C. Baldassarre, L. Gridinoc, M. Sabou, S. Angeletou, and E. Motta. Watson: Supporting next generation semantic web applications. 2007.

[2] M. D'Aquin and M. C. Suárez-Figueroa. A Quick Guide to Knowledge Reuse with the Watson Plugin for the NeOn Toolkit, 2008. `http://watson.kmi.open.ac.uk/DownloadsAndPublications_files/WNTP-guide.pdf`.

[3] P. Haase, H. Lewen, R. Studer, D. Tran, M. Erdmann, M. d'Aquin, and E. Motta. The neon ontology engineering toolkit. *In WWW*, 2008.

[4] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space: Theory and Technology*, volume 1. Morgan & Claypool Publishers, 2011.

[5] B. Hyland, G. Atemezing, and B. Villazon-Terrazas. Best Practices for Publishing Linked Data. W3C Working Group Note, 2014. `http://www.w3.org/TR/ld-bp/`.

[6] K. Janowicz, P. Hitzler, B. Adams, D. Kolas, and C. Vardeman II. Five stars of linked data vocabulary use. *Semantic Web*, 5(3):173–176, 2014.

[7] E. Jiménez-Ruiz, B. C. Grau, U. Sattler, T. Schneider, and R. Berlanga. *Safe and economic re-use of ontologies: A logic-based methodology and tool support.* Springer, 2008.

[8] D. Lonsdale, D. W. Embley, Y. Ding, L. Xu, and M. Hepp. Reusing ontologies and language components for ontology generation . *Data & Knowledge Engineering*, 69(4):318 – 330, 2010. Including Special Section: 12th International Conference on Applications of Natural Language to Information Systems (NLDB07) - Three selected and extended papers.

[9] M.-C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi. *Ontology Engineering in a Networked World*. Springer, Berlin, 2012.

[10] P.-Y. Vandenbussche, G. A. Atemezing, M. Poveda-Villalónc, and B. Vatant. LOV: a gateway to reusable semantic vocabularies on the Web. *Semantic Web Journal (under review)*, 2015. `http://www.semantic-web-journal.net/system/files/swj974.pdf`.

[12] `http://lov.okfn.org/dataset/lov/`

[13] `http://lov.okfn.org/vocab/voaf`

[14] `http://lov.okfn.org/dataset/lov/terms`

[15] `http://lov.okfn.org/dataset/lov/apidoc/`