DATALIFT: A PLATFORM FOR INTEGRATING BIG AND LINKED DATA

Gabriel Kepeklian, Laurent Bihanic

ATOS
Bezons, France

Raphaël Troncy

EURECOM
Sophia Antipolis, France

ABSTRACT

In the space domain, all scientific and technological developments are accompanied by a growth of the number of data sources. More specifically, the world of observation knows this very strong acceleration and the demand for information processing follows the same pace. To meet this demand, the problems associated with non-interoperability of data must be efficiently resolved upstream and without loss of information. We advocate the use of linked data technologies to integrate heterogeneous and schema-less data that we aim to publish in the 5 stars scale in order to foster their re-use. By proposing the 5 stars data model, Tim Berners-Lee drew the perfect roadmap for the production of high quality linked data. In this paper, we present a technological framework that allows to go from raw, scattered and heterogeneous data to structured data with a well-defined and agreed upon semantics, interlinked with other dataset for their common objects.

Index Terms— Datalift, Linked Data, Ontology, LOV, Geography

1. INTRODUCTION

The project Datalift, launched in late 2010 and supported by the French National Research Agency (ANR), has designed and developed a technical platform that aims to process raw data and to convert into semantified and interlinked data. It brought together eight partners coming from academia (INRIA, University of Montpelier, EURECOM) and industry (Atos, Mondeca) as well as two important French government agencies (IGN and INSEE) and the FING association for dissemination and outreach.

The Datalift platform realizes a virtuous circle in the processing of heterogeneous data. Every time it makes new data interoperable, its value increases. Indeed, the more data is linked, the more it gains value benefiting from the network effect. The platform takes as input raw data such as spreadsheet data (csv), geographical data (shp), structured data (xml) or even a connection to any relational database. In the space domain, the data is often described in XML (e.g. the Sentinel program).

XML is a universal meta-language that provides a consistent framework for the exchange of data and metadata between applications. However, XML does not provide any means to express neither the semantics (meaning) of the data, nor reasoning mechanism. For example, nested tags have no intended meaning associated: it is up to each application to interpret a nested structure. By contrast, linked data relies on ontologies, formal models that encode the intended semantics and vocabularies used by the data. Datalift is an open source platform that include numerous converters, for transforming raw data into RDF, the resource description framework model which is the basis of linked data technologies.

2. ONTOLOGY SELECTION

RDF provides a standard way to express simple statements about resources, using named properties and values. However, this models needs also to be complemented with a schema expressing constraints regarding the properties that can be attached to resources and their types: what types are allowed for a resources? what properties are allowed for a given resource type? what are the possible values for a given property? what are the relationships between resource types (generalization / specialization)? The role of ontologies is to cover this aspect: the formalization of vocabulary used in the data.

The publisher of a dataset should be able to select the vocabularies that are the most suitable to describe the data, and the least possible terms should be created specifically for a dataset publication task. The Linked Open Vocabularies (LOV) developed in Datalift provides easy access methods to this ecosystem of vocabularies, and in particular by making explicit the ways they link to each other and providing metrics on how they are used in the linked data cloud. LOV is integrated as module in the DataLift platform to assist the ontology selection.

3. DATA TRANSFORMATION

Once a suitable ontology has been selected, Datalift proposes an iterative process to transform the raw data into RDF while being compliant with the target model. The raw data is first converted into an initial RDF model following the W3C Direct Mapping conventions that were designed for mapping

SQL data into RDF but is here applied to any type of data source (CSV, XML, etc.). The conversion tries to preserve all the metadata available in the raw data (column names of CSV files, data types and relations (foreign keys) in SQL data, etc.). Once the data converted into RDF, a set of RDF-to-RDF transformation modules allow the user to incrementally transform the data to match the structures defined by each selected ontology.

Once the transformation process is completed, the data is ready to be made publicly available. This last step requires to define a URI naming policy to ensure data has consistent and and permanent identifiers. Datalift provides guidelines for defining such policies as well as tools to implement them, for example to separate resource URIs from representation URLs, perform content negotiation for RDF and non-RDF MIME types, etc. Finally, the publication step simply entails copying the final data from Datalift's internal RDF store into a public one. If a dataset catalog (DCAT) is available, Datalift will also populate it with a description of the published data.

4. INTERLINKING DATA

Just as data alone has little value, isolated RDF data would provide little added-value beyond compliance with a shared model, the business domain ontologies. To unleash the full power of RDF, data has to be linked with other referenced data. This is the interconnection step. Datalift provides tools to search for relationships between local data and existing public RDF data (i.e. open data available within remote SPARQL endpoints) and enrich the local data with links. Interconnection can occur at the dataset level (i.e. by comparing data) or at the ontology level (links between ontologies related to the same business domain that could easily be converted into links between data using these ontologies).

5. EXPLOITATION AND APPLICATIONS

While Datalift automatically exposes RDF data over HTTP, enforcing URI policies and content negotiation rules, it also provides additional tools to ease access to and consumption of RDF data, the most important of them being the SPARQL endpoint. Datalift SPARQL endpoint is compliant with the SPARQL 1.1 syntax, limited to read-only access, and includes support for RDF-based access control based on S4AC (Social Semantic SPARQL Security for Access Control): two users running the same SPARQL query will get different results, depending on the RDF graphs they are each allowed to access.

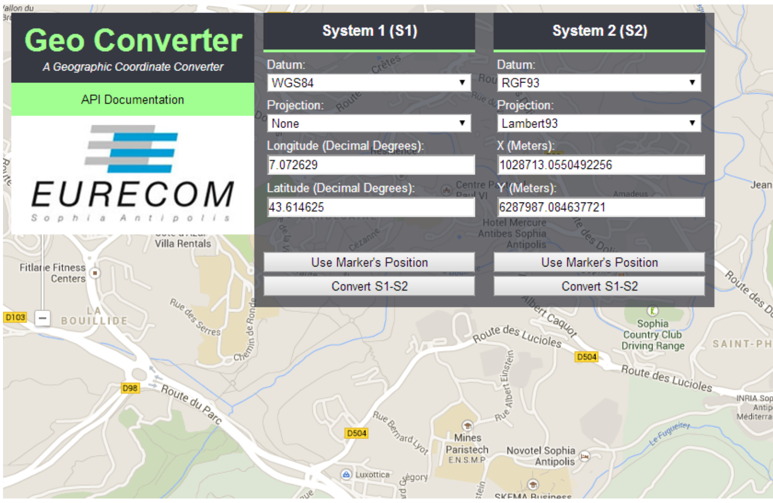
Other data access tools provided by the Datalift platform include SVG-based data visualization, HTML pages generation with RDFa tags, Graph Store HTTP Protocol support (in order to download a dump of an RDF dataset), etc. Datalift is a general purpose platform with which we have developed a number of use cases. INSEE and IGN have produced linked data that reference each other. In statistical datasets of INSEE, we can find `owl:sameAs` links to IGN geographical concepts, and conversely.

5.1. A Geo-converter for the Web of Data

For many years now, the web of data has been dominated with the use of only one Coordinate System (CRS), namely WGS84, to represent the localization of geographic objects on Earth. Nowadays, with the Open Data movement, more and more publishers including governments and local authorities are releasing legacy data that is often geolocalized in a different coordinate system. For example, IGN in France in releasing data that is geolocalized using Lambert93, a Lambert conformal conic projection (LCC) when objects are localized on the France metropolitan area. We have developed two semantic web vocabularies that take into account geometries defined in different coordinate systems and a REST web service that supports the conversion of coordinates between several CRS (Figure 1). The purpose of the REST Converter is to propose a web based service to perform conversion between various CRSs. The algorithms implemented are the ones described at `http://geodesie.ign.fr/index.php?page=algorithmes` and available within the standalone Circ software. At the moment, the following features are implemented in the Geo Converter:

- from/to WGS 84 to/from WGS 84 UTM ;
- from/to WGS 84 to/from Lambert 93 and
- from/to WGS 84 UTM to/from Lambert 93

The API can also convert a file with space separated values. The API supports JSON as one of the output format. The code of the REST service is available at `https://github.com/vienlam/Geo`

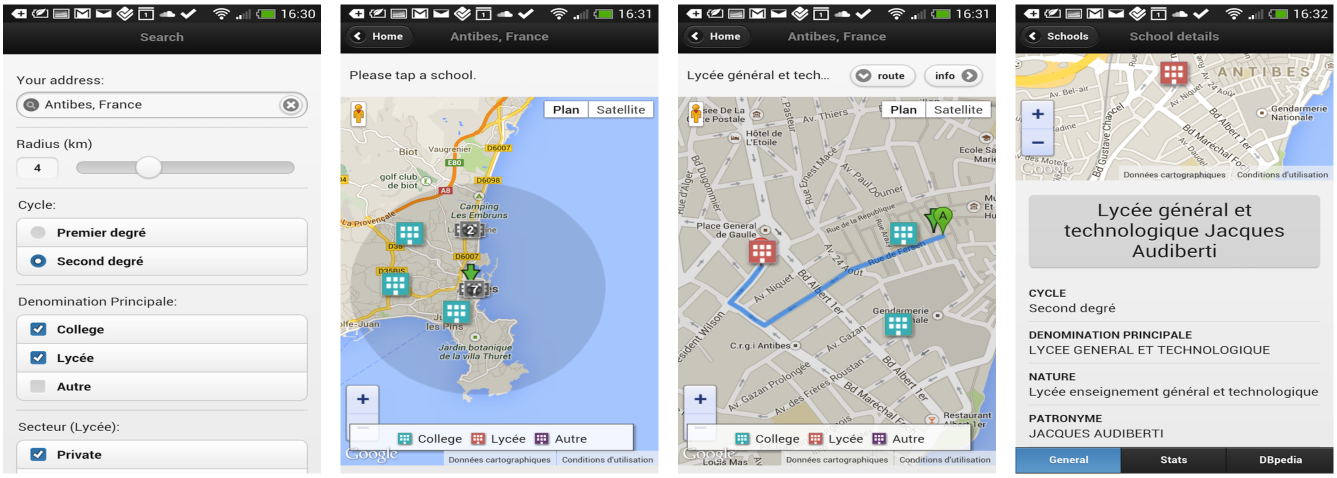


Geo Converter
A Geographic Coordinate Converter

System 1 (S1)
Datum: WGS84
Projection: None
Longitude (Decimal Degrees): 7.072629
Latitude (Decimal Degrees): 43.614625

System 2 (S2)
Datum: RGF93
Projection: Lambert93
X (Meters): 1028713.0550492256
Y (Meters): 6287987.084637721

Fig. 1. The User Interface of the Geo Converter



Your address: Antibes, France

Radius (km) 4

Cycle:
○ Premier degré
● Second degré

Denomination Principale:
☑ College
☑ Lycée
☐ Autre

Secteur (Lycée):
☑ Private

(a) Search options

Please tap a school.

(b) Results displaying on a map

Lycée général et tech...

(c) Route to the selected school

Lycée général et technologique Jacques Audiberti

CYCLE
Second degré

DENOMINATION PRINCIPALE
LYCEE GENERAL ET TECHNOLOGIQUE

NATURE
Lycée enseignement général et technologique

PATRONYME
JACQUES AUDIBERTI

(d) Details information with 3 tabs: General, statistics and DBpedia

Fig. 2. The PerfectSchool application developed with Datalift

5.2. Perfect School: an example of an application developed by Datalift

The Perfect School application is an example of application developed with the Datalift platform. This application aims to provide useful information on schools in France using semantic technologies, with RDF-ized data enriched with other datasets in the wild. The application and the vocabulary have successfully passed the integrity checker of an implementation for the candidate recommendation of Data Cube[1] vocabulary standardized by W3C.

In order to build such an application, we had to look at some relevant datasets we selected in the French government data portal `data.gouv.fr`. The ones selected for building the application are:

- `http://www.data.gouv.fr/DataSet/564055` in CSV format, containing a list of 67, 201 schools (name, status, type), with geolocation position in Lambert 93, for the academic year 2011-2012.
- `http://www.data.gouv.fr/DataSet/572165` in CSV format, giving indicators results for professional schoolsfor the 2011-2012 academic year.
- `http://www.data.gouv.fr/DataSet/572162` in CSV format, containing statistics for 2296 public high schools and indicators. It complements the statistics from INSEE.

We use the Datalift platform for transforming the different CSV files into RDF. We have also reused some external ontologies for ensuring interoperability:

- `aiiso`[2] for the type of school and codes of school.
- `geofla`[3] since the schools are considered as topographic entities.
- `geom`[4] for representing the different geometries (points with latitude and longitude) in a given coordinate reference systems with the `ignf` ontology at `http://data.ign.fr/def/ignf#`.
- `skos:Concept` for describing the 30 types of nature of schools.
- `qb:DimensionProperty`[5] and `qb:MeasureProperty` for modeling the dimensions and different indicators available for a given school.

The resulting vocabulary is available at `http://purl.org/ontology/dvia/ecole`. We also define different URI patterns for identifying all data objects. For example, The school "Albert Camus" in the city "Le Mans" with the code school 0720800D can we viewed in the application directly at `http://semantics.eurecom.fr/datalift/PerfectSchool/#school/0720800d/`

For the interconnection process, we used the Silk platform as it is re-packaged in the workflow of Datalift. Two datasets were used for finding `owl:sameAs` links:

1. DBpedia French chapter[6], as the scope of the application was limited to France. We have found only 7 match links with our schools datasets.

[1] `http://www.w3.org/TR/vocab-data-cube/`

[2] `http://vocab.org/aiiso/schema`

[3] `http://data.ign.fr/def/geofla#`

[4] `http://data.ign.fr/def/geometrie#`

[5] `http://www.w3.org/TR/vocab-data-cube/`

[6] `http://fr.dbpedia.org/sparql`

2. LinkedGeoData[7], as the underlying data used comes from the community project Open Street Map (OSM). Here, we got a total of 601 matching links in the category of `lgdo:BuildingSchool`.

The target device for the application is mobile phone, using principally two frameworks: Jquery mobile[8] and BackboneJS [9]. The application provides geolocation, search by city/district, graph charts for stats, table views of relevant results aggregated or group by some other aspects. The Perfect School Application provides 3 main views (Figure 2):

1. **Search form**: The interface retrieves the user location and offers choices based on the School type: first degree / second degree. When choosing first degree, the user can further select one of (primary school, elementary school or other). For the second degree, apart from looking for one of college, high school or other, the user can look for public or private schools. The search button launches the query for retrieving the collection of data matching the user's criteria.

2. **Search results**: The search action returns a collection of schools plotted on a map. A cursor on the left side helps users to zoom in and out to get more details about schools retrieved in a given area of interest. When selecting a given school, the name is displayed and with the possibility to see the route from the barycenter of the result on the map.

3. **Description of the school**: It is divided in 3 different tabs (a) General information (name, cycle, principal denomination, nature and patronym used); (b) Stats with all the different statistics in form of charts, graphs comparing the school with the others; and (c) DBpedia-FR[10] information if available, obtained with the `owl:sameAs` links for enriching the original dataset with information such as founder, date of creation, web site, population, head of school etc.

6. CONCLUSION

We have presented the Datalift platform for publishing and interlinking datasets on the web of Linked Data. We have described the framework architecture and its modules that support the selection of ontologies for representing the semantics of the data, the conversion of the data into RDF for a wide range of raw data formats, and the interlinking to other RDF datasets. While the Datalift platform has been used with early data publishers adopters such as INSEE and IGN in France, it remains a prototype tool. Its development is now ongoing as an open source project under the guidance of the Datalift association.

We have also described a generic geo converter web service that enables to transform a position on earth from one coordinate system to another. Finally, we have presented the PerfectSchool mobile application as an example of application that was built using the different modules of the Datalift platform. We will continue developing other applications that make use of the lifting process performed by the platform and show the added value of combining datasets to reveal new insights on the data. In particular we are working with with local and national government agencies to show how the publication and interlinking of datasets enable the development of innovative application for citizens.

7. REFERENCES

[1] Tim Berners-Lee, Chris Bizer, and Tom Heath, "Linked data-the story so far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.

[2] Tom Heath and Chris Bizer, *Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology*, Morgan & Claypool, 2011.

[3] F. Scharffe, Zhengjie Z. Fan, A. Ferrara, H. Khrouf, and A. Nikolov, "Methods for automated dataset interlinking," Datalift Deliverable D4.1, HAL, 2011.

[4] F. Hamdi, *Améliorer l'interopérabilité sémantique : Applicabilité et utilité de l'alignement d'ontologies*, Ph.D. thesis, Universit Paris-Sud XI, 2011.

[5] F. Scharffe, R. Troncy, G. Atemezing, F. Gandon, S. Villata, B. Bucher, F. Hamdi, L. Bihanic, G. Kepeklian, F. Cotton, J. Euzenat, PY. Vandenbussche, and B. Vatant, "Enabling linked-data publication with the datalift platform," in *26th Conference on Artificial Intelligence (AAAI)*, Toronto, Canada, 2012.

[6] H. Khrouf, V. Milicic, and R. Troncy, "Mining Events Connections on the Social Web: Real-Time Instance Matching and Data Analysis in EventMedia," *Journal of Web Semantics*, vol. 24, no. 1, pp. 3–10, 2014.

[7] R. Troncy, G. A. Atemezing, and N. Abadie, "Modeling geometry and reference systems on the web of data," in *W3C and OGC International Workshop on Linking Geospatial Data*, London, UK, 2014.

[7] `http://linkedgeodata.org/sparql`

[8] `http://jquerymobile.com/`

[9] `http://backbonejs.org/`

[10] `http://fr.dbpedia.org`

doi: 10.2788/1823