Publishing Linked Data Requires More than Just Using a Tool

G. Atemezing[1], F. Gandon[2], G. Kepeklian[3], F. Scharffe[4], R. Troncy[1], B. Vatant[5], S. Villata[2]
[1] EURECOM, [2] Inria, [3] Atos Origin, [4] LIRMM, [5] Mondeca

Abstract

Open Data raise problems of heterogeneity due to the various adopted data formats and metadata schema descriptions. These problems may be overcome by using Semantic Web technologies in order to move from raw data to semantic data interlinked in the Web of Data. However, lifting Open Data to Linked Open Data is far from being straightforward. In this paper, we describe the challenges we faced in developing the DataLift[1] platform, and the difficulties we encountered dealing with Open Data towards the publication of semantic interlinked data.

Introduction

As many initiatives around the world provide access to raw public data along the Open Data movement, many questions arise concerning the accessibility of these data. Various data formats, duplicate identifiers, heterogeneous metadata schema descriptions, and diverse means to access or query the data exist. These factors make it difficult for consumers to reuse and integrate data sources to develop innovative applications. Structured data is already present in databases, in metadata attached to medias, and in millions of spreadsheets created everyday across the world. 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. However, even if the raw data is there, even if the publishing and interlinking technology is there, the transition from raw published data to interlinked semantic data still needs to be done. We present Datalift, an open source platform helping to lift raw data sources to semantic interlinked data sources.

The ambition of DataLift is to act as a catalyst for the emergence of the Web of Data by providing a complete path from raw data to fully interlinked, identified, and qualified linked datasets. The Datalift platform supports the following stages in lifting the data:

1. Selection of ontologies for publishing data;
2. Conversion of data to the appropriate format (e.g., from CSV to RDF);
3. Interlinking of data with other data sources;
4. Publication of linked data;
5. Access control and licence management.

The remainder of the paper is as follows. First, we detail the main functionalities of the DataLift platform and its implementation. Second, we present the use cases we deal with and the problems that Open Data bring us when “translating” into semantic interlinked data. Finally, we conclude giving some future perspectives.

[1] http://www.datalift.org

Functionalities of the DataLift platform

The architecture of DataLift is modular. Several levels of abstraction allow decoupling between the different stages from raw data to semantic data. The dataset selection allows us to identify the data to be published and migrate them to a first RDF version. The ontologies selection step asks the user to input a set of vocabularies' terms that will be used to describe the lifted data. Once the terms are selected, they can be mapped to the raw RDF and then converted to properly formatted RDF. The data is then published on the DataLift SPARQL endpoint. Finally, the process aims at providing links from the newly published data to other datasets already published as Linked Data on the Web.

Dataset Selection

The first step of the data lifting process is to identify and access the datasets to be processed. A dataset is either a file or the result of a query to retrieve data from a datastore. The kinds of files currently considered are CSV, RDF, XML, GML and Shape files. Queries are SQL queries sent to an RDBMS or SPARQL queries on a triple store.

Ontologies Selection

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[2] (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 targets both vocabulary users and vocabulary managers: *i)* vocabulary users are provided with a global view of available vocabularies, complete with metadata enabling them to select the best available vocabularies for describing their data, and assess the reliability of their publishers and publication process, *ii)* vocabulary managers are provided with feedback on the usability of what they maintain and publish, and tools to show the dependencies and history of the vocabularies. LOV is integrated as module in the DataLift platform to assist the ontology selection.

Data Conversion

Once URIs are created and a set of vocabulary terms able to represent the data is selected, it is time to convert the source dataset into a more precise RDF representation. Many tools exist to convert various structured data sources to RDF[3]. The major source of structured data on the Web comes from spreadsheets, relational databases and XML files. We propose a two steps approach. First, a conversion from the source format to raw RDF is performed. Second, a conversion of the raw RDF into "well-formed" RDF using selected vocabularies is performed using SPARQL Construct queries. Most tools provide spreadsheet conversion to CSV, and CSV to RDF is straightforward, each line becoming a resource, and columns becoming RDF

[2] http://lov.okfn.org/dataset/lov/

[3] http://www.w3.org/wiki/ConverterToRdf

properties. The W3C RDB2RDF WG[4] proposes the Direct Mapping to automatically generate RDF from the tables but without using any vocabulary, and R2RML to assign vocabulary terms to the database schema. In the case of XML, a generic XSLT transformation is performed to produce RDF from a wide range of XML documents. The DataLift platform provides a graphical interface to help mapping the data to selected vocabulary terms.

Data Protection

This module is linked to Apache Shiro for obtaining the information, *i.e.*, username and password, about the user who is accessing the platform. The module[5] checks which are the data targeted by the user's query and then verifies whether the user can access the requested data. This verification leads to three kinds of possible answers, depending on the access privileges of the user: some of the requested data is returned, all the requested data is returned, or no data is returned. This means that the user's query is filtered in such a way that she is allowed to access only the data she is granted access to. The access policies are expressed using RDF and SPARQL 1.1 Semantic Web languages thus providing a completely standard way of expressing and enforcing access control rules.

Data Interlinking

The interlinking step provides means to link datasets published through the Datalift platform with other datasets available on the Web of Data. Technically, the module helps to find equivalence links in the form of "owl:sameAs" relations. An analysis of the vocabulary terms used by the published data set and a potential data set to be interlinked is performed. When the vocabulary terms are different, the module checks if alignments between the terms used by the two data sets are available. We use the alignment server provided with the Alignment API[6] for that purpose. We translate the correspondences found into SPARQL graph patterns and transformation functions are combined into a SILK[7] script.

Data Publication

This module aims at publishing the data obtained from the previous steps to a triple store, either public or private. The providers can restrict which graphs can be accessible, they could decide whether to provide just a "Linked Data" or a "Linked Open Data". Datalift comes by default with Sesame, but provides API for connecting to Allegrograph and Virtuoso triple stores as well.

[4] http://www.w3.org/2001/sw/rdb2rdf/

[5] http://wimmics.inria.fr/projects/shi3ld/

[6] http://alignapi.gforge.inria.fr/

[7] https://www.assembla.com/wiki/show/silk/

Lessons Learnt from the DataliftCamp

Datalift as a « *ready-to-use* » platform was tested in the release 0.67 during two days by providers of datasets, public authorities or enterprises willing to know how Linked Data can help them in their "day-to-day" business. It was a challenging task because for most of the 75 participants, it was the first time they dealt with terms such as "*vocabulary*", "*interlinking*", "*Linked Data*", "*OWL*", "*SPARQL*", etc. However, most of them were committed to follow the Open Data movement in France lead by *data.gouv.fr* and other initiatives from regions or cities.
Participants were grouped by domains according to their interests. As the platform can be easily installed on different OS (Windows, Linux, Mac)[8], it was easy for the participants to have a local copy installed for testing. Each group was assigned with one or two tutors with sufficient knowledge of the platform to guide them. Table 1 gives an overview of the different domains used to classify the providers, along with the provenance of the datasets, their formats and the 4-5 stars datasets in the LOD cloud identified for interlinking.

Table 1 - Datasets and scenario examples

Domain	Illustrative scenario	Original dataset	Original format	Datasets for interlinking
Person	Find translations of given names to different languages.	Opendata.paris.fr	CSV	DBpedia
Tourism, Culture and Events	Lift events data in the region Picardy	Yellow pages, Regional Tourism Office	XLS	EventMedia
Transport	Lift stops and bus transports	De Lijn bus company in Flanders (Belgium)	CSV GTFS[9]	--
Data Catalogues	Transform data catalogues using DCAT and thesaurus like EUROVOC for terms classification	Opendata.gouv.fr, Opendata.montpelliernumerique.fr	XLS CSV	--
Budget of collectivities	Transform using DQ, use SPARQL aggregation functions	Rennes, Montpellier, Toulouse	XLS	rdf.insee.fr
Geolocation	Interlinking,publishing different shape files with temporal data, geocatalog.	OpenstreetMap France Temporal series of agricultural data (confidential data)	SHP CSV	data.ign.fr rdf.insee.fr
Environment	Lift data of grapes of given parcel	Suez Environment INRA	CSV	DBpedia,rdf.insee.fr data.ign.fr

Going through the LD lifecycle with Datalift is not straightforward if we consider users that are not familiar to semantic slang and technologies. Providers had to face recurrent issues such as:

[8] http://www.datalift.org/en/node/24

[9] https://developers.google.com/transit/gtfs/

- *Choice of the suitable vocabulary that best covers the original dataset to lift*: Here it requires time and efforts to figure out which are the vocabularies in LOV where the terms can be reused without the need of creating new vocabularies.
- *Automatic detection of datasets to link to*, or how to go beyond the "default DBpedia" dataset for interlinking. Publishers may want to have a list of possible candidates of datasets to interlink to w.r.t. their own datasets.
- *Complexity of CONSTRUCT queries* that serve as an alternative to make RDF2RDF transformation in the actual version of the platform.
- *Time required for pre-processing tasks such as data cleaning* or normalizing all the attributes of a column field before using the first module of converting data to RDF.

Few datasets were finally published during the two days of the camp, although a big step was achieved in letting know to the different providers what is possible to achieve with Datalift for publishing their raw data as Linked Data. The camp leads to publish the catalog of the city of Montpellier in RDF using DCAT[10], and Open Food Facts data in RDF[11] and the food vocabulary[12]. We list here some recommendations that could help improving such tools like Datalift:

- Hide the complexity of SPARQL with natural language QA systems like QAKiS[13].
- Integrate a categorised list of candidate datasets worth to consider for linkage.
- Need of vocabularies integrating multilingualism to ease search using terms not from English.
- Need of tools for transforming Shape files to RDF according to any given geographic vocabulary and/or requirement.

Conclusions

Such experience with providers to lift their datasets was risky but beneficial. It was not at all easy for them to deal both with the use of the platform, and the key concepts of Semantic Web. The objective of lifting all the available datasets was not achieved and generated some frustrations from those who did not reach the final step. However we reached the important goal of generating interest and leading producers to ask questions about their data silo and the publishing process. This convinced ourselves about the expectations of the overall Datalift project. In the case of Open Data in France, we hope Datalift will contribute to say "*A little data lifted goes a long way*".

[10] http://opendata.montpelliernumerique.fr/datastore/VilleMTP_MTP_Opendata_2011.zip

[11] http://datahub.io/dataset/open-food-facts

[12] http://data.lirmm.fr/ontologies/food#

[13] http://dbpedia-test.inria.fr/qakis/