Llorenç Vaquer et al., Int. J. of Design & Nature and Ecodynamics. Vol. 11, No. 3 (2016) 370–375

DEVELOPING A DATA MAP FOR OPENING PUBLIC SECTOR INFORMATION

LLORENÇ VAQUER, JOSE VICENTE CARCEL, ANDRÉS FUSTER, IRENE GARRIGÓS, FRANCISCO MACIÁ, JOSE-NORBERTO MAZÓN& JOSE JACOBOZUBCOFF
University of Alicante, Alicante, Spain.

ABSTRACT

One of the key issues in developing an open data portal is detecting, selecting, classifying and prioritizing the data of the organization to be opened. It is not a trivial task since organizations (as universities) are rather complex (in terms of staff and existing information systems) and furthermore, data from universities has a rather heterogeneous nature. This paper introduces the "data map" concept as a means to support carrying out this process. A data map aims to help in improving the process of opening data, reaching a high level of automation by considering the required metadata and semantics.

Keywords: API restful, data map, metadata, ontology, open data, semantic web.

1 INTRODUCTION

Since the beginning of the digital information era, IT professionals have invested almost all their efforts to manage data in private scenarios because data were rather considered unshareable to obtain the most value. Nowadays, things have changed, and everybody considers that the data value gets maximized when it becomes open, i.e., freely available to everyone to easily reuse (http://opendatahandbook.org/guide/en/what-is-open-data/). Although private organizations are only starting to be aware of the importance of opening data, public institutions have understood that opening public data would produce a great benefit for the whole society [1]. Moreover, open data are a source of raw material and creativity for entrepreneurs and startups that can reuse data for creating value-added services through innovative business models [2]. According to the "Characterization Study of the infomediary Sector" conducted by the National Observatory for Telecommunications and the Information Society [3], the business volume directly associated with the activity of companies that reuse open data to generate applications products or services for third parties, ranges from 1.000 and 1.200 million euros. Other international reports [4] argue that the economic benefits (both, direct and indirect) of reusing open data in Europe is estimated about 200 billion euros annually (1.7% of the European GDP).

Currently, there are open data initiatives aligned with new political movements about Open Governments [5] based on transparency, participation and collaboration with the aim of sharing their public sector information. This open data movement is supported by different legislation and policies in Europe (Directive 2013/37 / EU of the European Parliament and of the Council of 26 June 2013 amending amending Directive 2003/98 / EC the reuse of public sector information (https://www.boe.es/diario_boe/txt.php?id=DOUE-L-2013-81251)and also at national level (Law 19/2013, of December 9, transparency, public access to information and good governance (https://www.boe.es/buscar/act.php?id=BOE-A-2013-12887)).

WIT CONFERENCES

This paper is part of the Proceedings of the International Conference on Big Data (Big Data 2016)
www.witconferences.com

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V11-N3-370-375

Therefore, public institutions around the world are making great efforts in opening data. There exist important tools like CKAN (http://ckan.org/) or Socrata (https://www.socrata.com/) designed to help organizations to publish and manage data, with the goal of becoming more open and transparent organizations for the society. Nevertheless, current approaches for opening data do not consider in a formal manner the current status of data sources, and how they may affect the opening process. Importantly, some pitfalls arise when opening public sector information, namely: (i) organizations originally store data with the aim of using it in their daily business (i.e., from a transactional point of view) without considering which are the best ways to store data for opening it and improving reuse; (ii) stakeholders of data sources may differ from those in the open data catalog, and they may thus have different features to be considered depending on the involved actor: data owner (entity that authorize or deny access to certain data), data responsible (person or people empowered for data management and data knowledge) or data consumer (entity, person or people that reuse and add value to the data). For example, there are cases in some open data catalogues that have some pitfalls that hinder the use of data. For example, (i) some catalogues provide data in files (CSV and XML) distributed over their website, but data consumers have to inefficiently visit different web pages and download several files to get the right data and metadata. Other problem that we find in some open data portals is the big amount of data available to download, this problem complicates the handle of data on the client side, sometimes making impossible to reuse. We have found these cases in data catalog from AEMET (http://www.aemet.es/ –Spanish Meteorological Agency) or Spanish Open Data Portal. They are open data catalogues, but the scenario nowadays required the metainformation to access data, therefore, mechanisms are required to know the state of data to facilitate opening and reusing.

In order to overcome these pitfalls, this paper aims at introducing the "data map" concept as a mean to support the process of selecting, classifying, and prioritizing the data of an organization to be opened. A data map aims to reach high level of automation when opening data by considering the required metadata and semantics directly from the data sources to be opened, thus ensuring open data quality [6]. Our approach is developed under the umbrella of the OpenData4U project from the University of Alicante (Spain) which aims to create a methodology to easily develop open data portals for universities. Actually, the open data portal of the University of Alicante (http://datos.ua.es) uses the approach presented in this paper.

2 OPENING DATA BY USING DATA MAP

Developing the Open Data Portal of the University of Alicante allows us to have enough experience to determine problems to be solved in the process of opening data [7]. One of the key issues is analyzing the current data sources in the organization before being opened. Within a public institution (as universities), there are many heterogeneous data sources with different features that must be known to be able to adequately open them. Interestingly, key issues in opening data of an organization are (i) finding the data sources within the organization (**where** are data stored?), (ii) determining kind of data sources, e.g. databases, unstructured files, etc. (**how** are data stored?), as well as (iii) detecting stakeholders which play an important role in the process of opening data (**who** is involved in storing data?).

- **Where are data stored?:** every kind of organization has a set of heterogeneous data source to open.Each data source depends on different departments and has different owners and responsible experts.
- **How are data stored?:**each data source includes rather heterogeneousfeatures due to the fact that access may be different (CSV, PDF, databases, web services, html). Data owner will know these features.

- **Who is involved in storing data?:**as we can see, data owner should have an important role during the opening process (data owner is also named data trustee, data custodian, data steward, data producer or data supplier in other areas). To ensure quality of published and maintain open data, data owner which is the person skilled in data and which will provide all the necessary information about data.

All the information collected from data owners are stored in a "data map" which contains information from data sources. Data map allows an organization to store these heterogeneous metadata in a homogeneous manner (see Fig. 1) as a previous step for storing data into a catalog to be opened. A data map helps us to organize information about data by using a structured schema. Information to be stored is: location data, data owner information (not only as a single person but as an entire unit, department, faculty or so on), access data information and legal information. A data catalog uses to be aligned with the Data Catalog Vocabulary (DCAT - https://www.w3.org/TR/vocab-dcat/).

Currently, as a first step, the data map is stored in a relational database with a simple schema as seen in Fig. 1. However, we plan to improve this data map developing an RDF vocabulary, so it is easier to add semantic information. In this way, the quality and possibilities of data will improve.

As a matter of fact, the "data map" bridges the gap, when publishing open data, between the data catalog, which contains the (open data) resources, and their corresponding metadata coming from the data sources (title, description, publisher or data owner, license, related data, categories, publication date, update date, update frequency, temporal coverage, some data quality criteria and the URL to download the data or resource). A data map can be used to increase the level of automation in publishing open data.

However, based on our own experience, we realized there was a problem with determining the meaning of the data to be reused: some items have an unrepresentative name, and they are not reusable without knowing the right metadata. Therefore, our data map is also concerned on collecting and storing the metadata straight from the data owner of the data sources to be easily included in the

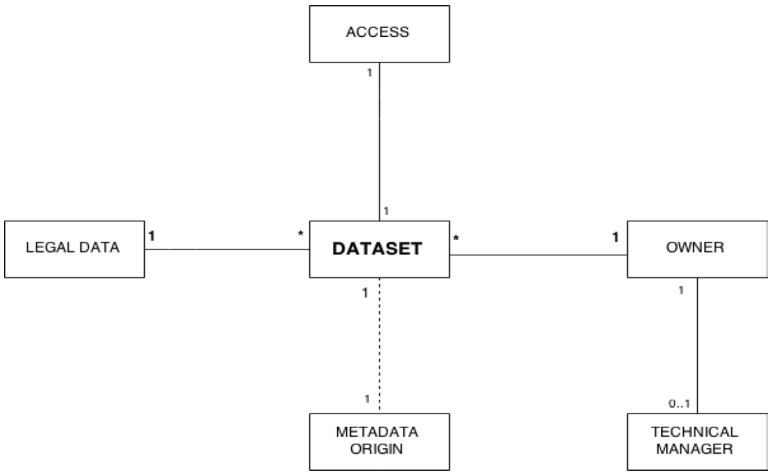

Figure 1: Schema of our data map.

data catalog. Metadata (containing the definition of the meaning of each item from data source) is then exposed together with the data.

Once the data catalog contains all the required information, the API Restful architecture provides a JSON structured format, taking into account any required format conversion (e.g. CSV to JSON). An overview of our approach is shown in Fig.2: our architecture is designed to collect all the information from a university into a data map (see Fig. 1). At the top of Fig.2, an ETL Bus is responsible for collecting the data regardless the source from which it comes (webservice, structured file, html, database, etc.) and fill the data map and data catalog. Once the data is in the catalog, it is accessible by using the API Restful (only requiring an API key).

Since open data aims to be reused by data consumers (e.g., developers), they need to access data quickly and easily. Importantly, the development of a data map allows us to provide the right information from the Open Data Portal (http://datos.ua.es/) at the right manner for developers (by deploying a RESTful API - https://dev.datos.ua.es/). This API has been already used in a contest in which several students participated by developing some applications (http://datos.ua.es/es/premios-concurso-aplicaciones.html).

Furthermore, this architecture is designed with decoupled modules that allow to be exported to any scenario. Of course, there is still much more work ahead, since our approach could be more powerful if it had a semantic component linked tothe data map, as stated in the next section.

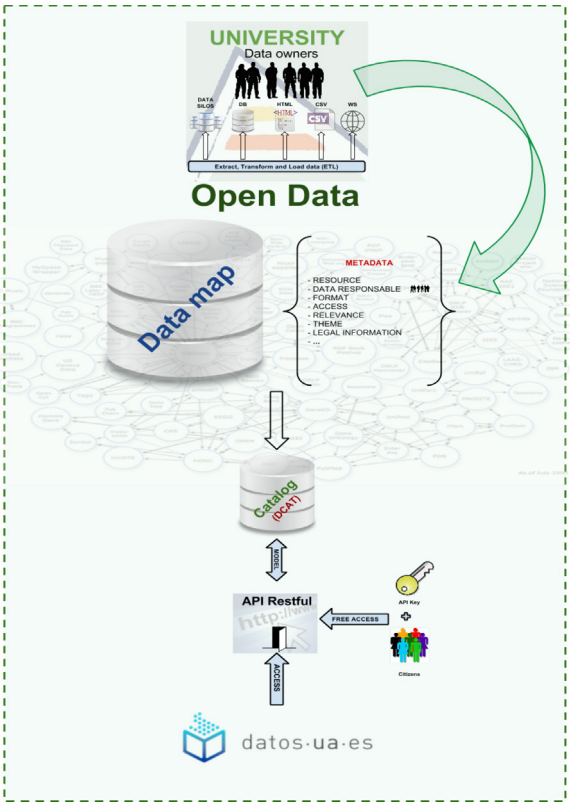

Figure 2: Open data portal architecture.

3 CURRENT CHALLENGES IN OPENING DATA: ADDING SEMANTICS TO DATA MAP

After the deployment of the open Data portal of the University of Alicante, data have started to be reused in applications and other services. Thanks to this, we realized about some issues who data consumers let us know: using our API was an easy and quick manner to access data but some effort must be done to process data to extract value. For example, to filter some data an entire dataset must be downloaded and parsed to obtain the right data. This process is time consuming at the client side since more information than required must be accessed.

Our first solution was offering a SPARQL Endpoint that allows data consumers to make queries directly to RDF graphs with a specific language. However, before taking this solution, we studied other initiatives which used the SPARQL Endpoint and the conclusions we obtained were not really good. The UK Government related their experience in [8]. They argue that RDF and SPARQL are new technologies for most developers and reusers, so the learning curve is very steep for developers. Even, worst, if the queries are a bit complex, they take too much time to be resolved. Authors then recommend a RESTful API to provide data. There is another solution proposed in [9] that provides a RESTful interface over the SPARQL technology, to make the way to access data easier, but unfortunately SPARQL underperformance is not solved.

Bearing these issues into consideration, some questions arise: How can we consider them our RESTful API? How can we get benefits from both SPARQL and RESTful? Our goal is having a RESTful API that can be queried with semantic information. i.e. we want to get the students enrolled this year in the career in architecture without having to download a data set containing all students enrolled in the university and then parsing it.

We envision an approach in which data consumers can access to the required information (specific data items) in a straightforward manner. To do so, datasets must be semantically tagged and linked, e.g., datasets must have the structure of an RDF graph. There exist some tools to convert structured files and relational databases to RDF (rdf123 - http://ebiquity.umbc.edu/project/html/id/82/RDF123, XLWrap - http://xlwrap.sourceforge.net/ or DB2RDF - https://www.w3.org/2001/sw/rdb2rdf/), but unfortunately they do not consider required metadata for opening, together with data.

Semantic tags can be used together with the metadata defined in our data map to create aRESTful API semantically enriched. Therefore, semantic resources related to the domain of the data sources are required. Regarding our example on universities, there are some vocabularies or ontologies thatcan be useful to enrich our data map, thus linking it with the organizational structure. In this way, open data would be linked to the supplier and all its metadata. For example, there are some ontologies created for academic institutions as AI- ISO (http://vocab.org/aiiso/schema) or Teaching Core Vocabulary Specification (http://linkedscience.org/teach/ns/), defined to implement and instantiate organizations like a university. However, those resources may lack in having every required definition, so they should be further studied. Once adding this semantic information, an enriched API Restful can be developed to provide every data about the university.

4 CONCLUSIONS AND FUTURE WORK

In conclusion, data map is the bridge to solve the metadata gap between data sources and data catalog. The information collected is crucial to determine the quality of data. This paper has focused on the solution adopted at the University of Alicante, this alternative can be applied to any context.

As a future work, we aim todevelop another RESTful API to provide open data as easy as possible, so every data consumer can reuse open data for designing their value-added products and services. Another task we plan to do is aligning our data map schema with DCAT vocabulary and

finding the best vocabularies for the domain to enrich data and to go toward automation the process to open data considering metadata and semantics.

REFERENCES

[1] Jetzek, T., Avital, M. & Bjorn-Andersen, N., The value of open government data: a strategic analysis framework. *In Proceedings of SIG eGovernment pre-ICIS Workshop*, Orlando, USA, 2012.

[2] Ferro, E. & Osella, M., Eight business model archetypes for PSI re-use. *In Open Data on the Web Workshop*, Google Campus:Shoreditch, London, 2013.

[3] ONTSI–ObservatorioNacional de lasTelecomunicaciones y de la Sociedad de la Información. Estudio de caracterizacióndel sector infomediario en España, 2014, http://datos.gob.es/content/estudios-de-caracterizacion-del-sector-infomediario-2014

[4] World Bank, Open data for economic growth, 2014, http://www.worldbank.org/content/dam/Worldbank/document/Open-Data-for-Economic-Growth.pdf

[5] Khatri, V. & Carol, V.B., Designing data governance. *ACM Digital Library*, **53**(1), pp. 148–452, 2010.
http://dx.doi.org/10.1145/1629175.1629210

[6] Oviedo, E., Mazón, J-N. & Zubcoff, J.J.,Towards a data quality model for open data portals. *CLEI*, pp.1–8, 2013.
http://dx.doi.org/10.1109/clei.2013.6670665

[7] Carcel, J.V., Fuster, A., Garrigós, I., Maciá, F., Mazón, J-N., Vaquer, L. & Zubcoff, J.J., Development of an Open Data Portal for a University - Experience from the University of Alicante. DATA, pp. 297–304, 2014.

[8] Sheridan, J. & Tennison, J., Linking UK Government Data. In *LDOW*, 2010.

[9] Coxa, S.J., Yua, J. & Rankineb, T., SISSVoc: A Linked Data API for SKOS vocabularies. Semantic Web Journal.