Converging on Semantics to Ensure Local Government Data Reuse

Laurens De Vocht¹, Mathias Van Compernolle², Anastasia Dimou¹, Pieter Colpaert¹, Ruben Verborgh¹, Erik Mannens¹, Peter Mechant², and Rik Van de Walle¹

> ¹ iMinds - Ghent University, Multimedia Lab ² iMinds - Ghent University, MICT Ghent, Belgium {firstname.lastname}@ugent.be

Abstract. When building reliable data-driven applications for local governments to interact with public servants or citizens, data publishers and consumers have to be sure that the applied data structure and schema definition are accurate and lead to reusable data. To understand the characteristics of reusable local government data, we motivate how the process of developing a semantically enriched exchange standard contributes to resolving this issue. This standard is used, for example, to describe contact information for public services which supports a representative pilot for opening up a variety of local government data. After implementing the pilot, we experienced that supporting the process of converging on semantics has a catalyzing effect on the reusability of government data.

1 Introduction

Linked Data arrived on the level of local governments, public services and their target user group: citizens. However, governments need to specify a desired reusable structure for their data before being able to use Linked Data. Governments take up this role to recommend describing data in the "Resource Description Framework" (RDF)³ and advocate the importance of this process. Initiatives such as "Open Standards for Linked (Government) Organizations" (OSLO)⁴ and the European Commission's "Interoperability Standards Agency" (ISA)⁵ enforce the use of Linked Data and its data model, RDF. Indeed, it is an interesting practice to tackle the semantic layer separately from the object, syntactic and application layer [11] within e-governments, which is made possible thanks to RDF.

RDF Similar data management tasks such as data alignment, data modeling or various data transformations are often repeated within a government. We can imagine the savings when the datasets are interlinked and reused when using RDF. For example, instead of making a *street event organization* application only for a single municipality, which outlines municipal services needed and permits required depending on the type of event, governments develop an event organization application usable for all municipalities in the region if they describe their required services needed to organize an event according to the converged semantics. Exchanging contact information, documents, reports and services benefit from such semantic convergence as well as all descriptions enriched with information annotated by other data publishers such as the datasets in the Linked Open Data Cloud.

⁵ http://ec.europa.eu/isa/

³ http://www.w3.org/RDF/

⁴ http://purl.org/oslo

If local governments would repeatedly develop an (ad-hoc) model for this information, it would be really hard for them to maintain the services that offer this information, as they require constant revising of the model of available data while they are not able to cope with newer technologies and applications without heavily investing in new support infrastructure. The opposite is also true, it is extremely easy to ignore this consideration and just publish data in whatever format pops-up at the time. Therefore, we ensure that after the standardization effort and as soon as the data has been described using one or more vocabularies that: the data still remains reusable, even when information technology evolves; and in one way or another, developers and users do not have to cope with the same issues as with custom data models.

Open standards are essential for systems and data to be interoperable [16]. Without open standards, open data can realize only a fraction of its value. Open data standards and architectures facilitate the integration with existing systems and allow data to at least be *complete*, *primary* and *timely*⁶. In the remainder of this paper we argue that the nature of the Linked Data published needs a useful integration of data across different administrations, where we aim at answering the question: *How and why is this integration effort for local and regional governments important for enterprises and citizens? How do open standards help in achieving this integration?*. We report on the standardization effort and how we extended it to apply to a pilot for local governments.

2 Related Work

Historically, governments tend to rely on the free market to disseminate information, e.g., the federal US Government, on the one hand, tries to reduce the role of the government in presenting its information to citizens [13]. In Europe, on the other hand, several frameworks showed up as a key tool for interoperability in the deployment of e-government services, both at national and at European level [8]. To untangle those legacies, many players in the private sector have embraced openness without reservation.

Methodologies for linking government data as such are not new: many guidelines considering applications, methodology, coverage and quality exist [19]. The Data-Gov Wiki project [6] is particularly relevant for linking government data. In this project data published at data.gov, a US portal for government data, was linked with LOD. It covers a range of topics such as government spending, environmental records, statistics on usage cost, use of public services. Around the same time, a similar effort for UK government data (data.gov.uk) emphasized why and how Linked Data was introduced and how a web of linked government data was created as part of the LOD cloud rather than focusing on linking [15]. These two government projects were preceeded by a debate on choosing for a closed or open warehousing model [18]. A tendency at the time of writing was data storage causing a high demand for metadata integration, which is in current terms translated to the need for a convergence on semantics of that data storage and at the time implied already the need for standardization.

Some initiatives succeed in this way of standardization, such as the *EDGAR* filings with '*eXtensible Business Reporting Language*' (XBRL)⁷: the 'Securities and Exchange Commission' (SEC)⁸ mandate requires that corporations and mutual funds under the purview of the Commission, file key performance reports in the Commission's EDGAR data repository⁹ in the XBRL format [4]. The Commission takes an approach to XBRL emphasizing use of high-quality vocabularies. Arguably, this does not solve

8 http://www.sec.gov/
9 http://www.sec.gov/edgar.shtml

⁶ http://opengovdata.org/

⁷http://www.xbrl.org/

the re-usability issue. There is no solution to this other than convincing governments to take on responsibility for providing an important public good [10]. For example, direct feedback such as an interoperability score when a new dataset has been created within a government can help the adoption of much used vocabularies [2]. Existing literature lacks concrete case studies on solutions for this issue, but the "Financial Information Observation System (FIOS)"¹⁰ uses Linked Data and multidimensional modeling of XBRL based on the RDF Data Cube Vocabulary¹¹ for accessing and representing relevant financial data [9]. Similarly, an RDF mapping could be defined for existing SOAP Web services [14]. The evolution towards open data contributes to the *timely* aspect because open data, either provided by the government itself or not, can enable the measurement of achieved objectives [12]. This helps citizens to understand and potentially allows for better competition.

3 Semantic Representations Supported by Open Standards

In an intergovernmental setting (with different structures as well as diverging public service models) it is complex to optimize the delivery of government services so that citizens and businesses only need to ask the government once for any of them. Even within the same organization, public services are documented following different flavors of national, regional or local public service models. Additionally, public service descriptions delivered through e-Government portals are usually unstructured and not machine-readable. This fragmented view of the public service concept and the absence of machine-readable public service descriptions impacts the quality and the efficiency of public service provision, increases administrative burdens and makes public service provision more costly. This is a major obstacle for citizens and businesses.

Smart Cities [1] are often confronted with: (1) a lack of open standards for (local) government data and (2) the vertical and cumbersome structure of their data architectures. The ISA core public services and the W3C Government Linked Data (GLD)¹² Working Group have provided standards and other information to assist governments around the world with publishing their data as effective and reusable Linked Data, using Semantic Web technologies. However, thanks to the flexibility that the RDF data model offers, local governments extend standards to serve their "custom needs" without diverging from the originals. The specification of the Flemish extension¹³ of OSLO¹⁴, was the result of our implementation of a public-private partnership. Our main task was to formalize the exchange standard and its extension.

In a typical scenario, local governments publish the open data, belonging to the deliverable for a project, as part of an application or report. After delivery, shortage of funding or change of focus to new projects leads to data which is not maintained. We assure that the published data can remain correct and up to date. Therefore, we modeled government data, described as Linked Data, with vocabularies in respect to their usage and wide popularity within the Semantic Web community as well as to their applicability for the proposed use case. The main modeling domains of interest as they were indicated by the local governments during the development of OSLO are: people, locations, organizations, products and services. To create new data sources, we aggregated information from different regional and local e-government information systems and combined them with existing services to create machine-readable public service descriptions. These descriptions are reusable, following the Linked Open Government

¹³ https://github.com/v-ict-or/oslo_xml_

schemas/tree/shared_catalogue_extension 14 https://github.com/v-ict-or/oslo_xml_

 $^{^{10}\; \}texttt{http://fios.ontologycentral.com/}$

¹¹ http://www.w3.org/TR/vocab-data-cube/

¹² http://www.w3.org/2011/gld

schemas/tree/master

Data paradigm [7] and enable functionalities such as automated service discovery and composition.

The Dublin Core vocabulary¹⁵ was used for the basic metadata properties, the Friend of A Friend (FOAF)¹⁶ ontology to bind the information on titles and descriptions and to enhance the content of the generated dataset. We used the *OSLO* vocabulary¹⁷ and the W3C Organization vocabularies: *Organization Ontology*¹⁸, *Registered Organization Ontology*¹⁹. For contact information we relied on the *VCARD Ontology*²⁰. We have extended the OSLO vocabulary with three new entities: *Channel, Activity* and *Product*. As a result we are able to model the latter three and *People, Organizations, Services* and *Locations*.

4 Shared Catalogue for Local Governments' Public Services

A number of administrations (local municipalities, cities and one regional administration) participated, as depicted in Table 1, in the development of a prototype for the "Shared Catalogue of Public Services", an ecosystem for sharing contact information from local governments: (i) towards citizens and (ii) between governments. The participating administrations all experienced a considerable "overhead" and "redundancy" in finding relevant contact data. The goal of this catalogue prototype is to disclose products and services more effectively between governments and towards citizens. Figure 1 shows how they interact using a common interface providing access to the data described semantically according to OSLO, enabling them to answer questions such as: (i)"As a *Citizen/Public Servant, how* to obtain *Service X* in *Municipality Y* and *who* to contact?"; (ii) "Who to contact for *Service X* in municipalities in a certain *region*?".

Administration	Туре	Conceptually Mapped	Formally Mapped	Automatically Linked
Gent	City	√	√*	√*
Kortrijk	City	√	✓	
Roeselare	City	~	✓	
Sint-Niklaas	City			√*
Beveren	Municipality			√*
Destelbergen	Municipality	√	✓	✓
Halle	Municipality	~		
Ingelmunster	Municipality	~	✓	
Knokke-Heist	Municipality	√	✓	
VDAB**	Regional	~		

 Table 1. Administrations participating in the Shared Catalogue and their mapping status.

 * In progress at the time of writing ** Flemish Employment Service

The pilot was initiated by *Corve*²¹, the Flemish e-Government coordination cell and the project consortium was coordinated by *V-ICT-OR*²², the Flemish ICT organization. We analyzed the contact data from several Flemish municipalities and cities and mapped according to the defined vocabularies. We dealt with multiple heterogeneous data that needed to be mapped. Each government had its own data structure. After mapping conceptually, i.e. matching concepts in the original datastructure with the OSLO vocabulary, we defined a formal mapping to RDF in a configuration document. This allowed us to align them for use in a common interface with one 'move'. A couple of governments, e.g., Beveren and Sint-Niklaas, could then be automatically linked without having to do the conceptual and formal mapping again, as they used a similar back-end data structure. It is noteworthy that for the biggest participating city, Gent, the mapping

20 http://www.w3.org/TR/vcard-rdf/

22 http://www.v-ict-or.be

¹⁵ http://dublincore.org/documents/dcmi-terms

¹⁶ http://xmlns.com/foaf/spec/ 20 http://www

¹⁷ http://purl.org/oslo/

¹⁸ http://www.w3.org/TR/vocab-org/

¹⁹ http://www.w3.org/TR/vocab-regorg/

²¹ http://www.corve.be

is more time consuming. This is expected as the delivery of product and services is more sophisticated and the data is more complex than for smaller governments.



Fig. 1. A shared catalogue from governments for citizens. A common interface provides a single access point thanks to vocabularies and open standards such as OSLO.

We relied on RML [5] to specify the mapping configuration. The data was based on publicly available website data, so that there were no privacy-related issues. Each municipality provided us with contact information of the members of the local government, their function and role in offered products and services for citizens (e.g., demand for a renovation permit or a new personal electronic passport). The mapped data was published using The DataTank [17], a system facilitating publishing datasets for governments. It has recently added the required support for the persistence of links and resolving Linked Data resources [3]. By applying this approach, publishers only need to worry about their source data and keep the links of published resources persistent, regardless of the used format – which are in most cases CSV files or relational tables.

5 Discussion and Conclusions

In our opinion, supporting data reuse implies preparing facilities in advance with the right vocabularies. This can be guaranteed on the ontology (semantic convergence) level and the data level. Support groups, mailing lists and open maintenance helped ensuring reuse of the ontology. In the case of data, and more specifically Linked Open Data, it is important to have a solid feedback loop so that the original data publishers can be notified at any point if there are any inconsistencies with the data they are responsible for. To avoid inconsistencies, we maximized automation between reliable authentic data sources and the published data.

In this pilot, we have experienced that the Linked Data principles are as suitable for data management on the Web as for local government information systems. We developed a standard as a convergence between various stakeholders in local government data. We showed its applicability using a prototype for a distributed shared catalogue of public services and products from municipalities. We are going to continue improving work in the field of Linked Open Government Data starting by assessing the user perceived usefulness and usability of this approach. On the one hand, this application is a typical use case for Linked Data technologies, given the complex and the dynamic nature of the organizations and data involved. On the other hand, it also takes place in a setting that is slow to evolve, partly for the same reasons. However, governments do have the potential to force or imply a certain structure using their own resources. Depending on the society model, short- and long term objectives, different technological choices need to be made.

We think that the Linked Open Data community should invest more in making its technology essentials clear and in motivating governments to use it as main technology for local government data reuse. It is an interesting research question to measure if and how well agreeing on semantics proves to be useful in tackling issues on converging semantics and reusable government data described in this paper.

References

- 1. Caragliu, A., Del Bo, C., Nijkamp, P.: Smart cities in Europe. Journal of urban technology 18(2), 65–82 (2011)
- Colpaert, P., Van Compernolle, M., De Vocht, L., Dimou, A., Vander Sande, M., Mechant, P., Verborgh, R., Mannens, E.: Quantifying the interoperability of open government datasets. Computer (Oct 2014)
- Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: Painless URI dereferencing using The DataTank. In: Poster and Demo Proceedings of the 11th Extended Semantic Web Conference (2014)
- Debreceny, R., Farewell, S., Piechocki, M., Felden, C., Gräning, A.: Does it add up? early evidence on the data quality of XBRL filings to the SEC. Journal of Accounting and Public Policy 29(3), 296 – 306 (2010)
- Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: a generic language for integrated RDF mappings of heterogeneous data. In: Proceedings of the 7th Workshop on Linked Data on the Web (LDOW2014), Seoul, Korea (2014)
- Ding, L., DiFranzo, D., Graves, A., Michaelis, J., Li, X., McGuinness, D.L., Hendler, J.: Data-gov wiki: Towards linking government data. (2010)
- Ding, L., Peristeras, V., Hausenblas, M.: Linked open government data. Intelligent Systems, IEEE 27(3), 11–15 (2012)
- Guijarro, L.: Semantic interoperability in eGovernment initiatives. Computer Standards & Interfaces 31(1), 174–180 (2009)
- Kämpgen, B., Weller, T., O'Riain, S., Weber, C., Harth, A.: Accepting the xbrl challenge with linked data for financial data integration. In: The Semantic Web: Trends and Challenges, pp. 595–610. Springer (2014)
- Leinemann, C., Schlottmann, F., Seese, D., Stuempert, T.: Automatic extraction and analysis of financial data from the EDGAR database. South African Journal of Information Management 3(2) (2001)
- 11. Melnik, S., Decker, S.: A layered approach to information modeling and interoperability on the Web. In: Proc. of the ECDL'00 Workshop on the Semantic Web (2000)
- 12. O'Reilly, T.: Open data and algorithmic regulation (2013)
- Robinson, D., Yu, H., Zeller, W.P., Felten, E.W.: Government data and the invisible hand. Yale JL & Tech. 11, 159 (2008)
- 14. Servant, F.P.: Linking enterprise data. In: Proceedings of the Workshop on Linked Data on the Web (2008)
- 15. Sheridan, J., Tennison, J.: Linking UK government data. In: Proceedings of the 3rd Workshop on Linked Data on the Web (2010)
- Simon, K.D.: The value of open standards and open-source software in government environments. IBM Systems Journal 44(2), 227–238 (2005)
- Vander Sande, M., Colpaert, P., Van Deursen, D., Mannens, E., Van de Walle, R.: The DataTank: an open data adapter with semantic output. In: 21st International Conference on World Wide Web, Proceedings (2012)
- Vetterli, T., Vaduva, A., Staudt, M.: Metadata standards for data warehousing: Open information model vs. common warehouse metadata. SIGMOD Rec. 29(3), 68–75 (Sep 2000)
- 19. Wood, D.: Linking government data. Springer (2011)