Supporting the vision of an open and interconnected administration using "Linked *Open* Data" principles

Enhancing data quality, accessibility and shareability

Fabian Cretton, Zhan Liu,
Anne Le Calvé

Business Information Systems

University of Applied Sciences Western Switzerland

Sierre, Switzerland

November 2013

HES-SO Valais-Wallis • rue de la Plaine 2 • 3960 Sierre
+41 27 606 89 11 • info@hevs.ch • www.hevs.ch

Valais excellence

Introduction

E-Government is a topical and necessary subject that includes cyber-administration and the equally important open data initiative. This report focuses on the basic strategy to allow the different actors publishing, accessing and manipulating data in order to fulfill any given task.

Here we use the term "actor" in a very broad sense: a citizen looking for information, an administration needing to share information with another administration, an official trying to understand the impact of a new law at different levels, and also a computer programmer who has to implement a new system or functionality. We thus consider the technical level, data publishers and consumers, business-to-business (B2B) as well as business-to-consumer (B2C) related issues.

Publishing data for humans is something well known, using written, visual or oral explanations in a specific human language. Making it available in a numeric format as PDF, image or sound files is also a mastered technology. However, problems arise when we are overwhelmed by data, and when the number of interlocutors or stakeholders increases. This is often the case in the E-Government domain, even more in Switzerland and its multilingual environment. As in other areas of our daily life, machines could be of great help if they could help us keeping track, finding, organizing and linking information, things our brain handles efficiently but only with a manageable amount of data and interlocutors.

In this regard, the emerging Semantic Web provides helpful technologies. Compared to traditional ways to organize data (files, RDBMs, APIs), those new technologies are conceived to share more easily data as well as their schemas, taking natively into account data quality and links between datasets. At the lowest level, they give machines comprehensive access to structured data, and highly promote good quality documentation for data consumers (developers for instance). At a higher level they thus allow the creation of powerful tools that can easily access and mix different sources of data, a hazardous task that is currently in the hand of the end-users.

5 stars data

The generalized approach when it comes to publishing machine-friendly data (i.e. structured data, not natural language) is to generate raw data in CSV files or spreadsheets, or in a more dynamic way through APIs. This is already good, but can be improved. Tim Berners Lee, inventor of the web and semantic web initiator, describes a 5 stars system[1] for open data:

[1] http://5stardata.info/

HES-SO Valais-Wallis • rue de la Plaine 2 • 3960 Sierre
+41 27 606 89 11 • info@hevs.ch • www.hevs.ch

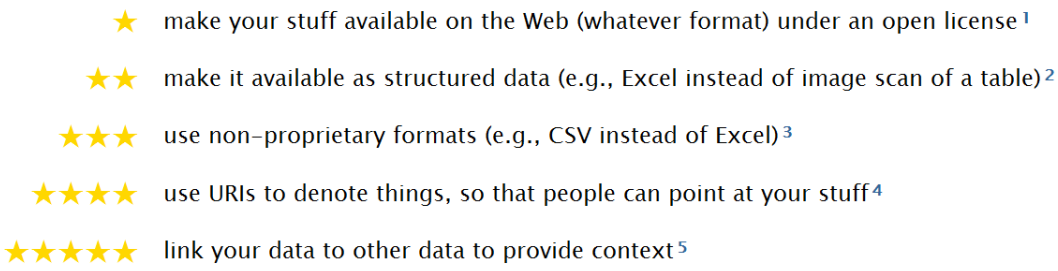

Figure 1 - 5 star deployment's scheme

The current standard to publish 5 stars data is called RDF (Resource Description Framework)[2], and the web of data that is referred to as Linked Open Data (LOD)[3]. The schemas (ontologies) are mostly designed using RDFs[4], SKOS[5] or OWL[6]. The schemas and the data form an RDF graph, which can then be queried using the SPARQL[7] language. All those components of the semantic web stack are W3C standards which are well implemented in the current tools.

Here is a comparison table explaining the possible improvements to the current situation:

Raw data/API	Semantic Web
Difficulty to find the data or the web services providing the data	Data is **in** the web, can be searched using appropriate tools
Each data source is published in its own format: a data consumer needs to understand separately each raw data file and API, which becomes much harder when the number of sources increase	All data are published in the common RDF format, and can thus be queried with a single language
The data schemas are more or less well documented and available	Schemas (ontologies) and data can be queried with the same language. Moreover, as the goal is to share a schema, it is a common practice to add human labels and comments to the ontology itself, and to publish (generate) a documentation web page[8]
Each piece of data might not have a universal identifier	Each piece of data (a resource) is identified by an URI, most often an URL in the web context
Semantic about the data (what it means, how to use it, etc.) is often found in the code using the data (thus not necessarily available to others)	Semantic about the data is self-contained and goes with the data
Datasets are not linked: a data consumer has to find the links between datasets, which is a time consuming task, provided it is possible	Publishers of data provide links as often as possible, see the LOD cloud [9] for instance. Creation of new links is easy.

[2] http://www.w3.org/TR/rdf-primer/
[3] http://linkeddata.org/
[4] http://www.w3.org/TR/rdf-schema/
[5] http://www.w3.org/2004/02/skos/
[6] http://www.w3.org/TR/owl2-overview/
[7] http://www.w3.org/TR/sparql11-overview/
[8] Example for the music ontology: http://musicontology.com/specification/
[9] http://lod-cloud.net/

HES-SO Valais-Wallis • rue de la Plaine 2 • 3960 Sierre
+41 27 606 89 11 • info@hevs.ch • www.hevs.ch

We see in the above table that some benefits come from the fact that information which is traditionally separated from the data (its schema, its semantic) can now be managed within the data itself.

Doing things in a "richer" way opens up new possibilities, but it does of course have some cost (also listed on the 5 star system web page). Here is an overview for a data provider:

- Convert the data to the common format and assign/manage URIs
- Instead of designing the data schema from scratch (which might be faster but not so clean in this context), time must be taken to analyze which vocabulary to reuse
- Instead of simply publishing raw data, effort is needed to find out how to link to other datasets

For a data consumer, dealing with more complete data and schemas opens up a world of new possibilities, but require more effort to understand the data model. We might say that there is less effort needed to access the data, but more effort to understand it in order to make a good use of it.

As we can see, the benefit of such an approach has to be evaluated globally. A data provider will put some effort to publish he's data, and he will be rewarded if the data is effectively found and used (especially if he provides data catalogues for instance). Search engines are already first class consumers of structured meta-data[10]. The data provider will then also benefit from other data sources published in the same way.

Linked Open Government Data (LOGD)

E-Government throughout the world is already experiencing with those technologies, especially when it comes to opening data. The W3C has a dedicated working group "Government Linked Data (GLD)"[11], and early this year a "Case study on how Linked Data is transforming E-Government"[12] was published by the European Commission.

The European Commission supports the ISA[13] program to facilitates interoperability, sharing and re-use between European Public Administrations. Its Action 1.1[14] is about "Improving semantic interoperability in European eGovernment systems" through the "Semantic Interoperability Community" known as SEMIC[15]. Their experience and guidelines would be useful for creating the Swiss LOGD.

[10] http://schema.org
[11] http://www.w3.org/2011/gld/wiki/Main_Page
[12] https://joinup.ec.europa.eu/community/semic/document/case-study-how-linked-data-transforming-egovernment
[13] ec.europa.eu/isa
[14] http://ec.europa.eu/isa/actions/01-trusted-information-exchange/1-1action_en.htm
[15] https://joinup.ec.europa.eu/community/semic/description

HES-SO Valais-Wallis • rue de la Plaine 2 • 3960 Sierre
+41 27 606 89 11 • info@hevs.ch • www.hevs.ch

Swiss LOGD data publishing

To maximize the efficiency and quality of a national LOGD, there is a need for a semantic web policy that helps the different actors to follow standards, design and maintain core vocabularies if needed, promote the re-use of the core vocabularies or other existing schemas, and thus facilitate datasets linkage.

This should help and promote good practice in order to:

- Design and share URI's
- Design and share common schemas or core schemas
- Publish schema documentation
- Define the way to publish correctly RDF schema and data, handling dereferenceable URI[16] and content negotiation[17]: this means that any URI can be accessed on the web, and that the service returns RDF or HTML depending on the consumer. See for instance http://dbpedia.org/resource/Switzerland which returns the same information but as http://dbpedia.org/page/Switzerland for HTML consumer (a browser) or http://dbpedia.org/data/Switzerland for RDF consumer
- Handle licensing and versioning
- Reference the published data in a common repository, in order to facilitate the collaboration of semantic web developers (or consumers) for the Swiss LOGD. For instance, how do one know which eCH recommendation has been published in RDF, to which Swiss government RDF data should one refer when publishing its own, what are the common URIs to use when referring to locations, organizations, people, etc.

Analyzing a use case

When taking part in the Linked Data approach, the primary question is not about "what will be the use of the data", but the focus is on providing quality data and links to other datasets. The use of the data is totally open to data consumer, and thus not restricted to the provider's point of view.

When a new actor wants to take part in the Swiss LOGD movement, he could follow those steps:

- Analyze and clarify which data he can provide, as well as the corresponding metadata
- List the form/source of the data (e. g. CSV, PDF, RDBMs, WebService), and see how to convert it to RDF
- Determine how he's own data is inter-related and how it can be linked. A single actor is maybe dealing internally with silos of data
- Determine to which other data from LOGD his data and metadata relates. He could even think more globally about any LOD data
- Determine to which other data source his data could relate if they were available as RDF
  This could push other actors to contribute to the Swiss LOGD or LOD: if one publishes its data, it might make sense if that other data is also available in RDF so that the datasets can be linked.

[16] http://en.wikipedia.org/wiki/Dereferenceable_Uniform_Resource_Identifier
[17] http://en.wikipedia.org/wiki/Content_negotiation

HES-SO Valais-Wallis • rue de la Plaine 2 • 3960 Sierre
+41 27 606 89 11 • info@hevs.ch • www.hevs.ch

This is maybe currently done by a specific software, by the end-user himself and manually, or maybe not even feasible.
- Then the above mentioned steps for "Swiss LOGD data publishing" should be followed

Further detailed information can be found in the W3C Linked Data Cookbook[18] or in the free guide "Linked Open Data: The Essentials - A Quick Start Guide for Decision Makers "[19].

Conclusion

In this presentation of Linked Open Government Data we discussed how semantic web technologies would enhance the current situation to publish Open Data and Linked Data, moving toward a 5 stars open and interconnected administration.

We thus provided:
- A comparison table of the current API and raw data situation with the one that could be achieved relying on semantic web technologies
- The cost of publishing data in a better way
- A proposal and guidelines to promote a Swiss LOGD initiative
- A description of the first steps that a newcomer could undertake to see if its use case does fit the LOD approach

To conclude we want to stress out that there is no need to make a clear choice between the possible technics. The linked data approach is above all a *state of mind* to share and reuse (data, schemas, etc.), before being a purely technical question. Taking a step toward semantic web technologies does not mean to turn the back on the current technologies as "APIs/XML/raw data" with which developers are currently more familiar. Both of them can co-exist, both have their flaws and advantages, and we are willing to support any useful approach.

[18] http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook
[19] http://www.semantic-web.at/LOD-TheEssentials.pdf

HES-SO Valais-Wallis • rue de la Plaine 2 • 3960 Sierre
+41 27 606 89 11 • info@hevs.ch • www.hevs.ch