FREIE UNIVERSITÄT BERLIN

How are People Engineering Linked Data?
A Survey Snapshot about the Engineering Efforts Spent by Dataset Publishers

Markus Luczak-Rösch

TR-B-11-09
December, 2011

FACHBEREICH MATHEMATIK UND INFORMATIK
SERIE B • INFORMATIK

1 Introduction

Linked Data is at the moment potentially the most successful notion of the Semantic Web with respect to public visibility and attention. At least the open government data efforts, which adopted the Linked Data paradigm prove that. Research of recent years has brought up a variety and also in practice well-understood set of principles and methods how to publish Linked Data on the Web quickly and easily. It is also a widely accepted insight that the understanding and adoption of Linked Data consumption and thus the uptake of real-world applications is far from being successful due to multiperspective issues with respect to dataset and data quality amongst other issues.

In this technical report we present a preliminary compilation and interpretation of the "LOD Provider Survey 2010". The motivation for this **qualitative survey** was to empirically answer the following two questions in the context of the Web of Linked Data:

1. *Which dedicated engineering processes are performed to create the ontologies which underly the datasets in the LOD cloud?*
2. *Do the dataset publishers have a structured process for dataset maintenance?*

Our understanding of ontology engineering refers to the definition by Gomez-Perez et al., given as follows:

> *"Ontological Engineering refers to the set of activities that concern the ontology development process, the ontology life cycle, the methods and methodologies for building ontologies, and the tool suites and languages that support them."*

We understand dataset maintenace in the context of this paper as follows:

Definition 1. ***Dataset maintenance*** *is the continous contribution of a dataset publisher to the overall goals of the Web of Data.*

These goals are best described by Bizer and Heath who distinguish between the classical data integration scenario, where the consumer has to bear the integration effort, and the Web of Data, where

> *"data publishers may contribute to* ***making the integration easier for data consumers*** *by reusing terms from widely used vocabularies, publishing mappings between terms from different vocabularies, and by setting RDF links pointing at related resources as well as at identifiers used by other data sources to refer to the same real-world entity."*

Given this goal and the high level principles to achieve it, it is reasonable to expect from research to come up with methods, tools, and studies that help to follow a set of principles and to understand the benefits of dataset maintenance. This will bring people in practice to the point to adapt this process step to the lifecycle of their respective datasets.

2 Why is Ontology Engineering Crucial for the Dataset Publisher?

That real ontology engineering is something a dataset publisher is confronted with is best described by an example from one of our previous studies on SPARQL query logs. Consider a user asking for the instruments which appear in the music of The Beatles, which is a reasonable question when we take into account that there is a dedicated wikipedia page on that[1] starting with the information "The Beatles started out like most other rock and roll bands, employing a standard guitars/bass/drums instrumentation. . . ". When we translate this into a series of SPARQL queries of a DBpedia user, the conflict between the published data on DBpedia conforming to the DBpedia ontology and the user's information need in the context of the property "instrument" comes clear. Figure 1 depicts this example.

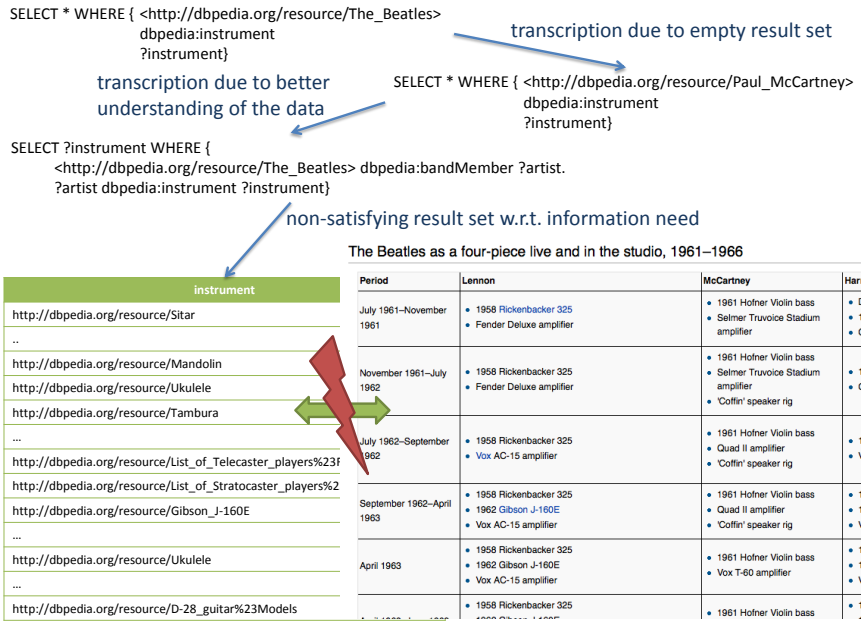

Fig. 1. Visualization of how the populated DBpedia data fails to fulfill a user's request for the instruments which were characteristic for The Beatles as a band and not the ones each single artist was able to play.

Some trivial resolution patterns to this issue could be to

[1] `http://en.wikipedia.org/wiki/List_of_The_Beatles\%27_instruments`

– simply populate the data that corresponds to the wikipedia page `http://en.wikipedia.org/wiki/List_of_The_Beatles\%27_instruments` and conforms to the triple pattern `SELECT*WHERE{<http://dbpedia.org/resource/The_Beatles>dbpedia:instrument?instrument}`
– change the domain property of dbpedia:instrument to be restricted to "MusicalArtists" and not "Thing", so a "Band" is explicitly invalid in this context

In fact there are many other ways of potential remodeling that could help to resolve the above mentioned conflict. We choose these two because they are representative for an interesting issue. The effect of the first resolution can be measured objectively as an increasing correspondance of a dataset to the requested queries of the users. But, the effect of the second resolution depends on subjective criteria, which make the dataset publisher to express that such a query, is out of scope of the dataset.

For this example we do not take into account if such a change results in any change on the technical infrastructure, e.g. algorithms to perform the population of instances conforming to the applied ontology.

3 Related Work

For this preliminary technical report we keep the survey related work relatively short. We admit that quality criteria and evaluation of datasets has recently been addressed recently [1]. This stresses that data quality and dataset evaluation is of actual interest. In the Web community a discussion about dataset quality criteria based on the data quality assesments presented by Pipino et al.[2] has taken off [3]. This directly relates to objective measures and subjective perceptions of data and consequently fits well in the context of our example.

It is a wide spread and reasonable perspective that the Web of Data is still the Web and thus the user has to respect the diversity, heterogenety and potentially inconsistency of data in this open information space. So it is also clear that efforts appear which run on as much of the published datasets as possible in order to help the data consumer to understand the structure, strength and weaknesses of this global repository [4,5]. With the prespective of an ongoing growth of the Web of Data this will become more and more complex at least in terms of computing measures on the whole amount of data, though the claim for effort distribution between the publisher and the consumer for quality assesing tasks is emphasized.

4 Design of the LOD Provider Survey

The survey was performed as a qualitative online survey within a period of four weeks starting at the beginning of October 2010. The questionnaire was not announced and released to the public but to a dedicated group of 100 distinct mailaddresses of people. These addresses were found during an online search in repositories that list the actual available interlinked datasets and give at least

a pointer to some project responsible person (e.g. CKAN and the LOD W3C SWEO Community Project[2]. The addressed people represented in total 216 datasets which is a coverage of 100% of the LOD cloud at that time conforming to the publicly announced and listed datasets. Emails explaining the survey and inviting the people to participate were send out. A first wave addressed all the collected addresses and the second one just a selected list of 12 people which we personally know and which did not answer the survey in reaction of the first wave. In the remainder of this section we describe the layout of the survey questionnaire.

Meta Questions About the Dataset and the Publisher: On the meta level we asked the participants for their name, affiliation, mailaddress, the name for the represented LOD dataset, and at least one public SPARQL endpoint of this dataset. At the end of the questionnaire the people were also able to add free text comments.

General Questions about the Ontologies Used: In order to find out information about the overall amount and the character of the ontologies used for data publication we asked the following two questions:

1. How many ontologies do you use to populate your dataset altogether?
2. Which are the ontologies you use? (give names or URIs)
3. How many of the used ontologies did you develop yourself?

In fact, the first two questions would be something we could also find out by analyzing the datasets in terms of namespace identifiers used and other structured information. But, since the major goal was to find out something about the the self-developed ontologies we decided also to ask for these two trivial facts in addition.

Questions Related to the self-developed Ontologies: For up to five ontologies we asked the following questions for each of the ontologies individually:

1. How did you develop your ontology?
   - manually from scratch
   - ontology reuse and manual adaption
   - ontology learning
   - automatic generation from any semi-structured datasources
   - automatically derived from any relational database
2. Did you follow any methodology for engineering the ontology and if yes, which one?
3. What is the size of this ontology in terms of the number of concepts?
   - Vocabulary (up to 150 concepts)
   - Small ontology (between 150 and 1000 concepts)

[2] `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

   - Mid-sized ontology (between 1000 and 5000 concepts)
   - Large ontology (more than 5000 concepts)
4. What is the complexity of these ontologies in terms of the usage of ontology language primitives?
   - RDF-S
   - OWL-Lite
   - OWL-DL
   - OWL-Full

Question related to the Evolution of the Ontologies: In order to capture information about the anticipation of dataset publishers about future evolutionary steps for the self-developed ontologies we asked again globally and not individually for each single ontology:

1. Do you see any need to evolve these ontologies in the future?
2. Why?/Why not?

5 Survey Results

26 participants filled out the questionnaire which is **a response rate of 26%** of all contacted people. Three persons mentioned several ontologies which were self-developed but only for a subset of these ontologies the details on the engineering were given properly. We eliminated this entries from the data. There was also one overlap in responsibility for a dataset, so a person which was originally addressed as the contact person for another dataset felt also responsible for DBpedia. We also eliminated the two DBpedia related responses from the data. A fourth entry had to be eliminated because it was mentioned that no ontology was self-developed however details on ontology engineering were given. After this we end up with **a population of 20 properly filled questionnaires**.

One person was representing 31 datasets with the same properties regarding the questionnaire which he noted as a comment and which was reproducible from the researched list of datasets and the project responsible people. This leads to an abolute number of 50 represented datasets which means that **this survey covered round 23% of all LOD cloud datasets which were publicly available in October 2010**.

Table 5 presents absolute and relative values compiled from answers to the genral question 1 and 3.

# of datasets covered	50
# of ontologies to populate data	65
average # of ontologies per dataset	1,3
# self-developed ontologies to populate data	21
average # of self-developed ontologies per dataset	0,42

How did you develop your ontology?	
manually from scratch	62% (13)
ontology reuse and manual adaption	0%
ontology learning	0%
automatic generation from any semi-structured datasources	5% (1)
automatically derived from any relational database	33% (7)

What is the size of this ontology in terms of the number of concepts?	
Vocabulary (up to 150 concepts)	86% (18)
Small ontology (between 150 and 1000 concepts)	9% (2)
Mid-sized ontology (between 1000 and 5000 concepts)	0%
Large ontology (more than 5000 concepts)	5% (1)

In the following we compile the results of the questions related to the self-developed ontologies which means that percentages relate to 21 as 100%.

That any engineering methodology was followed is mentioned in 20% of all responses (4 times). 3 times (15%) "OntoClean" is mentioned as the methodology which was performed and 1 time (5%) the participant relates to the Linked Data principles as the methodology.

The question whether the participants see any need to evolve the ontologies in the future is positively answered in 15 times (75%). Every single participant who did this also gave a comment on why this is expected. The compilation of potential reasons for this is shown in Table 5. One participant who gave a negative answer to this question gave a reason against any need for evolution as follows: "They are standard ontologies".

6 Discussion and Concluding Remarks

It is our preliminary interpretation of the survey that there is a missing understanding of a structured maintenance plan when publishing data and associated ontologies. We conclude this from the followig results of our survey:

– In 80% of all cases the dataset publisher does not follow any methodology to develop the ontologies underlying her data, thus there is no lifecycle model that requires at least a critical review of the vocabularies used for data publication at some point of time.

What is the complexity of these ontologies in terms of the usage of ontology language primitives?	
RDF-S	52% (11)
OWL-Lite	0%
OWL-DL	48% (10)
OWL-Full	0%

Why do people expect that evolution is necessary?
"More vocabulary mappings" (was given 3 times)
"Mappings to more ontologies. Now only mappings to some others." (was given 3 times)
"Changes based on recommendations and discussions by various standards groups."
"In the future I may need to add new terms into the ontologies. The ontologies were constructed to support the linked data being used and as that data evolves new predicates and classes made need to be added to the ontology."
"These ontologies are used by other datasets and a wider community - they will need to adapt to its evolving needs."
"Because deriving predicate names from databases schema works, but actually modeling the real life concepts in each database would be more useful."
"YAGO is constantly being improved and expanded."
"Move properties and classes will need to be included if more data from other sources is added."
"simplification, discussions in community, change in scope, etc."
"We have more data we want to expose"
"Data-driven need for new properties and classes."

– The Linked Data principles have been regarded as an ontology engineering methodology. Since these design issues do not recommend or propose any kind of a dedicated lifecycle model people are not stimulated to regard data publishing as a continuous process.

Since 75% of the respondents stated that they see the need to evolve their ontologies in the future we conclude that there is an awareness for a dedicated ontology and dataset lifecycle. The interviewees also give very concrete reasons for their assessment in this case, which is a nice starting point for methods and tools aiming to provide the dataset publisher with maintenance activities.

That 86% of the self-developed ontologies are vocabularies with at most 150 concepts and that 52% remain on the level of RDF-S expressivity indicates that ontologies underlying LOD are rather small and simply structured.

6.1 Selfcritical Discussion about the Methodology of the Survey

As it was already mentioned before, the survey was ran in the end of 2010. Today it is possible to recognize a slight change in the perception of the lifecycle of linked datasets. This becomes evident at least by publications in the filed of life cycles for data centric systems such as [6] which is accepted but not yet published officially. But still, since this is only a high level survey, real methods, tools and studies are missing which is a barrier for the adoption of dataset maintenance as a common task.

One commentary on this survey was that OWL 2 profiles are missing for the classification of the complexity of the ontologies. We note this for a possible

second round of this survey. In the current one all participants classified their ontologies conforming to one of the given answer possibilities.

The survey focused dataset maintenance from the ontological perspective and not the pure technical perspective. That means that it focuses on criteria which have nothing to do with any process and technical infrastructure that helps to re-perform the publishing process iteratively as it is very well performed and documented for the DBpedia project for example. Such criteria are about the knowledge that is provided by the data as a result of the underlying conceptualizations and how the maintenance of this dimension is planned and performed. We also explicitly respect that some projects – and again DBpedia is one of the most prominent example and Freebase is another one – provide an infrastructure that enables the community to edit the schema. However, the tools provided by these projects do not enable any kind of requirements analysis and consensus finding processes which yield in the end to an ontology change operation.

This preliminary study was conducted in order to found the motivation that there is a need for research on dataset maintenance from the perspective of the dataset publisher. It was also meant to be a starting point to collect information about an evolution in this area. As a byprodcut it gives an insight on the limited adoption of existing ontology engineering methodologies in the context of the Web of Data. This inspires to study in detail whether this is logical or a result of a misconception of ontology engineering.

References

1. Hoxha, J., Rula, A., Ell, B.: Towards green linked data. In: Proceedings of the Second International Workshop on Consuming Linked Data (COLD2011), CEUR Workshop Proceedings (CEUR-WS.org) (Oktober 2011) vision paper.
2. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Commun. ACM **45** (April 2002) 211–218
3. Hartig, O., Flemming, A.: Quality criteria for linked data sources. Website (2011) Available online at `http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Quality_Criteria_for_Linked_Data_sources`; visited on 2011-12-19.
4. Guéret, C., Groth, P., van Harmelen, F., Schlobach, S.: Finding the Achilles Heel of the Web of Data: using network analysis for link-recommendation. In: International Semantic Web Conference 2010. (forthcoming 2010)
5. Guéret, C., Groth, P., Stadler, C., Lehmann, J.: Linked data quality assessment through network analysis. (2011)
6. Möller, K.: Lifecycle models of data-centric systems and domains. Accepted paper at the Semantic Web Journal, waiting for publication (2011) Available online at `http://www.semantic-web-journal.net/content/lifecycle-models-data-centric-systems-and-domains`; visited on 2011-12-19.