universität bonn

Rheinische
Friedrich-Wilhelms-
Universität Bonn

Institut für Informatik III
Abteilung Enterprise
Information Systems

universität **bonn** · Informatik III · Römerstr. 164 · 53117 Bonn

Prof. Dr. rer. nat. Dipl. Math.
Sören Auer
Leiter der Abteilung Enterprise
Information Systems

Postanschrift:
Römerstr. 164
53117 Bonn

Tel.: +49 (228) 73-7816
Fax: +49 (228) 73-4382
auer@cs.uni-bonn.de
http://eis.uni-bonn.de

Aktenzeichen:

Bonn, 25.03.2015

Thesis review „Publishing and Consuming Government Linked Data on the Semantic Web“ by Auguste Ghislain ATEMEZING

In the last years, the domain of Open Government Data received increasing attention. Within governments and public administrations the awareness grew, that opening and sharing of government data has a number of advantages and benefits for citizens, companies, governments, the public sector itself as well as society at large. For citizens, for example, the availability of open government data can significantly reduce the barrier between governments and the general public. Citizens can engage with their governments through the analysis of data, the proposal for planning and new directions of policy making and governance. For companies, the abundant availability of open government data can result in the creation of new data-driven products, services and business models. Unfortunately, we learned that the publishing of data by governments and public administrations is not enough. The plethora of formats (e.g. XML, CSV, Excel, PDF or even HTML) the lack of nomenclature, standards and semantics as well as licensing issues (e.g. non-conformance with the open definition) hinder the realization of the benefits of open government data. Using semantic web standards and technologies can help to address the most pressing problem in this context, the variety, heterogeneity and diversity of data formats, structures and representations. This is the context of Auguste Ghislain Atemezing's PhD thesis.

In the first chapter of his thesis, Auguste Ghislain Atemezing presents the context of the thesis. He introduces the Linked Data concept as well as the domain of geographic information (which is a key data domain for open government data). He derives a number of research questions, which are broadly covering various technical areas of geospatial open data publication and consumption. Finally, the chapter summarizes the contributions of the thesis and presents an outline through the thesis structure. The challenges are well illustrated in this chapter, but my impression is that the thesis covers a too broad area. From my point f view, it would have been better to focus a bit more. Since most of the

aspects covered in the thesis are related to spatial data, the title of the thesis should reflect this better. A more suitable title, for example, could be “Publishing and Consuming Geo-Spatial Government Data on the Semantic Web”.
Chapter 2 gives an introduction to geospatial data on the Web. The representation of spatial data is described and illustrated as well as the formats and serializations. The second section in this chapter presents a REST service for converting geo data. A question here would be why content negotiation can not be applied in addition to encoding the result formats in URL parameters. The author introduces a number of spatial data representations and vocabularies. He nicely discusses requirements and compares features of these representation formalisms. He proposed and implemented three vocabularies for geometries, coordinate reference systems (CRS) and topographic entities. The vocabularies extend existing vocabularies and add two additional advantages: an explicit use of CRS identified by URIs for geometry, and the ability to describe structured geometries in RDF.

Chapter 3 tackles the publishing, interlinking and querying of geodata. After introducing a number of existing tools, the author identifies their limitations. Criteria for interlinking are presented and functions for geometry comparisons (Hausdorff, Euclidean distance) discussed. The next section in this chapter compares the features of a number of triple stores with particular emphasis on geospatial features. With GeoKnow Workbench and the Datalift platform two comprehensive frameworks are discussed. In addition, the final section highlights the publishing of French authoritative data using the latter framework. Again, a number of aspects, such as data conversion, URI design and interlinking are nicely discussed and illustrated with examples. The two pages on evaluation of spatial queries in the end of the chapter address this problem rather superficially. Here a more detailed analysis would have been desirable.

Chapter 4 analyses existing tools for the consumption of Linked Data using visualizations and applications. The chapter reviews generic visualization tools as well as Linked Data specific ones. A survey table juxtaposes the tools in terms of data formats, libraries, licenses and languages. An interesting contribution here is the vocabulary for Describing VIsualization Applications (DVIA), but again a more detailed presentation, evaluation and application of the vocabulary would have been desirable.

Chapter 5 is devoted to the creation of visual applications. A key contribution of this chapter is the LDVizWiz: a Linked Data Visualization Wizard. The corresponding section describes the workflow and the individual steps. While a formalization of the workflow as given below figure 5.1 is a very good idea, the actual definitions and formalizations are not properly introduced and described. For example, it is unclear why the sets of vocabularies and domains have different cardinality (m and n) and the function $\Phi(L,D)$ is not defined.

Chapter 6 describes contribution to the Linked Open Vocabularies infrastructure. It presents the how LOV can be used with an ontology methodology (e.g., the NeOn methodology) to improve reuse of vocabularies. The author proposes and implements a heuristic to align vocabularies on the Web of Data and a ranking of vocabularies based on information content (IC) metrics. Again, this could have been easily a

topic of a thesis on its own and is rather superficially covered. Then, a module of Datalift that reuses the LOV catalogue in the process of converting raw dataset for reusing terms already defined in other vocabularies is described.

Chapter 7 tackles the problem of license compatibility checking, while the work is concluded in chapter 8 with an outlook on future work.

Overall, the thesis describes a substantial amount of engineering work. My main criticism is that the work is too broad and some aspects are covered only superficially. A more limited and focused work with clearer defined research questions would have represented a stronger contribution. Also, I suggest to focus the work a bit more in the title. There are very substantial areas of open government data publication and consumption, which are not addressed by the thesis at all or only superficially – such as, for example, statistical data (in 5.6, which is one of the largest governmental data domains. A clearer focus of the thesis title on spatial government data is from my perspective advisable. Another weakness of the thesis is that it is not completely clear, what was the work of the author and of other collaborators. When looking at the publication list, there is a large number of co-authors, some of which might have used parts of the work in other PhD thesis. This is fine, but should be made more explicit, for example, by adding a footnote to every chapter or key section indicating the publication this particular section is based on and the approximate quantitative and/or qualitative share of the author. The publication list is impressive, but most of the publications are workshop or demo articles.

The thesis is well written, edited and illustrated. There are a few inconsistencies in capitalization and punctuation (e.g. spaces before punctuation marks).

Given the overall high engineering-quality of the thesis and the comprised technical and application work being performed by the author, I can recommend the thesis for the defence.

I'm looking forward to attend the thesis defence soon.

Best regards,

Prof. Dr. Sören Auer