Request for GeoNovum Tender

Wouter Beek (wouter@triply.cc) and Laurens Rietveld (laurens@triply.cc)

Triply (http://triply.cc)

1 Specification of the task for which we are applying

Triply applies for 'Research Topic #1 Revisited: Modern Ways of Spatial Data Publication, Task 1 (Data Publication)'. The research question of this topic is: "How do these lessons learned meet the constraints (e.g. budgets) and capabilities (e.g. in-house know-how) of governmental organizations on the one hand, and of data users on the other?"

2 Motivation for applying for this research topic

Triply applies for this research topic because it sees innovation potential for the way geodata is currently published on the Web of Data. Triply wants to explore and concretely deliver significant improvements to the way in which geodata is disseminated to data consumers. Triply believes that it is possible to unlock innovation potential within this tender given (1) the close collaboration with Geonovum, (2) the presense of other data suppliers such as PDOK and cultural heritage organizations, and (3) the lessons learned during the previous tender round on which we can build.

When we look at contemporary geodata publishing practices we see two worlds colliding: On the one hand there are data publishers that use OGC standards that are very expressive but also very complex. On the other hand there is the Web of heterogeneous and distributed data sources. There is a large number of (potential) data consumers that are using the contemporary Web stack and are used to far simpler and more Open Web APIs. (These are the two parties that are also identified in the research question for this research task.)

Luckily, usability and complexity need not exclude each another once the Linked Data approach is applied correctly. Since Linked Data does not require a single database schema to operate, we have the ability to disseminate the same data through different but complementary vocabularies. For instance, there can be two schemas: one complex and difficult and antoher simplistic and easy (lesson 1B). In this example the two vocabularies cater towards the aforementioned two user groups. In this way the same data can be

delivered in different ways, benefitting different groups (lesson 1). While the specification of multiple vocabularies/schemas on top of the same data collection comes at a low cost it does provide the necessary tuning to user capabilities (lesson 1D).

The other part of the research question targets the "constraints (budgets) and capabilities (in-house know-how)" at the data publisher side. Operating costs can be split into fixed and variable costs. The **fixed cost** of deploying a Linked Data solution (e.g. a remotely accessible triple store) is considerable. At the moment governmental organizations wish to use and/or provide a Linked Data deployment they oftentimes have in-house people that do a considerable portion of the engineering. There is no clear reason why Linked Data deployment should require more in-house knowledge then using the electric grid. With the Triply Linked Data product a (governmental) organization uses the Triply Linked Data infrastructure in a similar way in which it uses the electric grid; it simply connects. Because Triply significantly reduces the fixed cost of deploying a Linked Data solution, the (governmental) organization can focus on curating the data and extracting value from it.

The **variable cost** of deploying a Linked Data solution is an entirely different story. While it is somewhat difficult to set up a triple store it is virtually impossible to build one that scales to service a large amount of (consecutive) users. When we look at the SPARQL endpoint observatory SPARQLes (`http://sparqles.ai.wu.ac.at`) [Buil-Aranda et al. 2013], we see that 80% of current SPARQL endpoints have low availability. When we look at the 20% SPARQL endpoints with high availability, we see that they enforce restrictions on the kinds of questions that can be asked and the size of answer sets that are returned. This effectively means that remote triple stores are either unailable or severely constrained. No Web programmer wants to program against a database with either of these properties. Triply is the first Linked Data solution that allows the number of (consecutive) users to scale without blowing up the variable cost of endpoint operation. Triply is able to do this because Triply is based on recent innocations in large-scale and scalable Linked Data deloyments: Linked Data Fragments (LDF) [Verborgh et al. 2014], Header Dictionary Triples (HDT) [Fernández et al. 2013] and LOD Laundromat [Beek et al. 2014].

3 Plan of approach

The plan of approach focuses on realizing maximal impact within a short period of time. All crucial decisions in the plan of approach will be made by Triply's CTO who has extensive experience with building large-scale data systems. Some of the low level tasks in the plan of approach will be performed by a Triply engineer.

3.1 Requirements analysis

Triply will collaborate with Geonovum, data providers and task 2 partners in order to make a requirements analysis that specifies what is needed in order to support the goals of the overarching tender call. Making the requirements analysis a shared responsibility will ensure that development of the demonstrator will be demand-driven.

3.2 Data conversion and cleaning

Since some data may either not yet be formatted as RDF or may not yet be fully standards-compliant, Triply will perform data conversion and cleaning steps in order to ensure that all data meets certain quality criteria. Inherent bugs in the data that cannot be fixed by technological means alone will be communicated back to the original data suppliers.

3.3 Database layer

The result of the data conversion and cleaning steps will be populated into a SotA database backend. The backend will be chosen so as to be able to service at least one thousand simultaneous API users. The focus is here on features that Web developers find beneficial, i.e., prioritizing scalability, efficiency and performance (lesson 1B).

3.4 API layer

Triply will define and implement a RESTful Web API that follows the lessons learned, together with formal and de facto standards in Web API construction. This RESTful API layer will focus on lowering the barrier for (Web) programmers to use Geodata. In line with lesson 3A, content negotiation will be used to let a data consumer specify the format s/he prefers.

3.5 Iteration cycle

Once the whole system is fully functional and has task 2 partners programming against it, Triply will assess which changes are needed in each of the layers in order to improve the demonstrator for maximal utility.

4 Results

This section enumerates the results that will be made available.

4.1 Data

Triply will make all cleaned, converted, linked and otherwise enriched data available to the original data suppliers. The data will contain future-proof persistent URIs (lesson 2C) and will contain structured annotations that make it easy to find when published online (lesson 2D). Persistent URIs will be based on the Dutch Linked Data URI Strategy Overbeek and Brink 2013.

4.2 Vocabulary

Triply will make the vocabulary that bridges the geodata and Web API communities available to Geonovum and/or the original data suppliers.

4.3 Demonstrator

Triply will provide a demonstrator Web Service that will be used by task 2 collaborators. Data in the demonstrator will be supplied in standardized Linked Open Data formats (lesson 3A).

4.4 Reports

Triply will write a report in which it will validate the lessons learned with respect to the practice of making large quantities of geodata available in a Web- and developer-friendly way.

4.5 Dissemination

Finally, Triply will communicate the results obtained within the Geonovum tender with the wider Linked Open Data, Geodata and Web API communities. Concretely, Triply will perform the following dissemination activities:

1. Triply will, in collaboration with partners, present the results of the Geonovum tender at community events about Linked Open Data, Geodata, and Geo-oriented Web APIs.
2. Triply will write a report describing the tasks that have been performed. The report will also contain the findings and answers to the task description questions (as enumerated in the "Invitation to tender" document).
3. Triply will take the lead in writing a white paper about the requirements for the deployment of a scalable Linked Geodata backend and an approachable Web API for Open Geodata. The paper will be written in close collaboration with the various partners and will be submitted to an appropriate venue (e.g., WWW, ESWC Industry Track or Semantics).

4.6 References

Triply is a startup that is based on research previously conducted by the Knowledge Representation and Reasoning (KR&R) group at VU University Amsterdam (VUA). Headed by Prof. Dr. Frank van Harmelen, this is the top research group in the world in the area of Semantic Web and Linked Data research.

The founders of Triply have proven their ability to **conduct outstanding research** in the field of Semantic Web and Linked Data. This has resulted in multiple highly cited research papers over the last two years alone. The Triply founder's research effort has culminating in winning the Best Research Paper Award at the International Semantic Web Conference (ISWC) 2015, which is the top venue for Semantic Web research. Here follows a selection of relevant publications over the last two years:

- Wouter Beek et al. (2014). "LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data." In: *The Semantic Web–ISWC 2014*. Springer, pp. 213–228

- Laurens Rietveld and Rinke Hoekstra (2014). "YASGUI: Feeling the Pulse of Linked Data." In: *Knowledge Engineering and Knowledge Management*. Springer International Publishing, pp. 441–452
- Wouter Beek and Laurens Rietveld (2015). "Frank: The LOD Cloud at Your Fingertips." In: *European Semantic Web Conference (ESWC) 2015: Developers Workshop*
- Laurens Rietveld et al. (2015). "Redeploying the Semantic Web: Linked Data as a Service." In: *ESWC 2015*
- Laurens Rietveld, Wouter Beek, and Stefan Schlobach (Under submission). "LOD Lab: Experiments at LOD Scale." In: *The Semantic Web–ISWC 2015*. Best Paper Award
- Jan Wielemaker et al. (2015). "ClioPatria: A SWI-Prolog Infrastructure for the Semantic Web." In: *Semantic Web Journal*
- Filip Ilievski et al. (2016). "LOTUS: Adaptive Text Search for Big Linked Data." In: *ESWC 2016*
- Wouter Beek et al. (2016). "LOD Laundromat: Why the Semantic Web Needs Centralization (Even If We Don't Like It)." In: *IEEE Internet Computing* 20 (2)
- Wouter Beek, Stefan Schlobach, and Frank Van Harmelen (2016). "A Contextualised Semantics for owl:sameAs." In: *ESWC 2016*

In addition to delivering outstanding research, the founders of Triply have proven their ability to **build outstanding Linked Data solutions** as well. Firstly, the LOD Laundromat (`http://lodlaundromat.org`) ecosystem has won the Best Dutch Linked Open Data Award 2015 (and ended 3^{rd} in the European competition). LOD Laundromat is widely recognized as a Game Changer within the Semantic Web research community. Secondly, Triplys CTO has build Yasgui (`http://yasgui.org`), the most feature-rich SPARQL editor. Yasgui is used by many high-impact Linked Data projects and parties such as Sesame, Jena, Smithsonian Museum, German National Library of Economics and Linked Open Vocabularies. Thirdly, Triplys CEO is involved in the development of the ClioPatria triple store (`http://cliopatria.swi-prolog.org`) and SWI-Prologs Semantic Web library (`http://www.swi-prolog.org`). Finally, the founders of Triply have recently been involved in the construction of LOTUS (`http://lotus.lodlaundromat.org/`), a large-scale and customizable Semantic Web search engine.

4.7 Indication of in-kind investment

Triply will match Geonovums investment amount with an in-kind investment of equal size, i.e., €10.000,- The in-kind investment is necessary to realize the ambitions expressed in the plan of approach.

4.8 Statement of agreement

Triply agrees to publish all research results and deliverables that flow from this tender under the CC/by license.

References

Beek, Wouter et al. (2014). "LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data." In: *The Semantic Web–ISWC 2014*. Springer, pp. 213–228.

Buil-Aranda, Carlos et al. (2013). "SPARQL Web-Querying Infrastructure: Ready for Action?" In: *The Semantic Web–ISWC 2013*. Ed. by Harith Alani et al. Vol. 8219. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 277–293.

Fernández, Javier D et al. (2013). "Binary RDF Representation for Publication and Exchange (HDT)." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 19, pp. 22–41.

Overbeek, Hans and Linda van den Brink (2013). *Towards a national URI-strategy for Linked Data of the Dutch public sector*. Tech. rep. KOOP and Geonovum.

Verborgh, Ruben et al. (2014). "Querying datasets on the Web with high availability." In: *The Semantic Web–ISWC 2014*. Springer, pp. 180–196.