What is the current state of the Multilingual Web of Data?

Asunción Gómez-Pérez and Daniel Vila-Suero
Ontology Engineering Group
Universidad Politécnica de Madrid
asun@fi.upm.es, dvila@fi.upm.es

1. Motivation

The Semantic Web is growing at a fast pace, recently boosted by the creation of the Linked Data initiative and principles. Methods, standards, techniques and the state of technology are becoming more mature and therefore are easing the task of publication and consumption of semantic information on the Web.

As identified in [1] this growing Semantic Web offers an excellent opportunity to build a multilingual "data network" where users can access to information regardless the natural language they speak or the natural language the information was originally published in. But it also creates new research challenges and presents some risks, being the most significant one the creation of, what the authors describe as, "monolingual islands" – where different monolingual datasets are disconnected to datasets in other languages.

Having this in mind, we pose ourselves two simple questions:

- *Are we able to devise representative statistics and findings that could help us to shed some light on the current state of the Web of Data with respect to the use of natural languages?*
- *Can such statistics and findings serve us in the development and testing of new tools, techniques and services that could help to overcome the aforementioned challenges?*

The preliminary work we present here represents an effort to gather useful information and resources that could help to face already identified research challenges, discover new ones and provide a base for discussion within the research in Multilingual Semantic Web.

Our initial objective will be to answer questions like:

- *What is the distribution of natural languages on the Web of Linked Data?*
- *To which extent are language tags used to indicate the language of property values?*
- *Which domains are predominantely mono/multilingual?*
- *What is the distribution of cross-lingual links vs. monolingual links?*
- *How are cross-lingual links established (e.g. owl:sameAs)?*
- *Are we able to identify monolingual datasets not connected to data in other languages and thus conforming "monolingual islands"?*
- *Do mono/multilingual datasets organize themselves into clusters with respect to the used natural languages?*

The remainder of the text gives an overview of the resources we count on in order to set up an environment for our study (Section 2), and the methodology of our study (Section 3).

2. Resources

With the increasing amount and heterogeneity of data being published to the so-called Web of Data, in the recent years we find (1) several measurements and empirical studies of web data, (2) crawled datasets providing representative subsets of the web of data.

Regarding (1), we find several works such as [2,3,4]. However, most of them do not take into account the multilinguality at all (e.g. [2,3]) or do it to a limited extent (e.g. [4]). In [4], Ell et al. introduce a set of label-related metrics and report their findings from measuring a subset of the Web of Data[i] using the proposed metrics. One of these metrics is the multilinguality of the labels. More recently we find the *LODStats*[ii] initiative that aims at gathering comprehensive statistics about datasets adhering to the RDF found at thedatahub.org. In the website we find statistics about languages[iii], however it remains unclear how these data are gathered and it lacks absolute numbers that could help to analyse for example the distribution of usage of language tags (i.e. to which extent are language tags used to indicate the language of the property values).

Regarding (2), since 2009, a number of crawled corpora are being made publicly available in order to facilitate the analysis and characterization of web data. One of the most significant examples is the Billion Triples Dataset, already used for a number of studies (e.g. [4,5]). More recently, we find the "Dynamic Linked Data Observatory"[iv], a framework to monitor Linked Data over an extended period of time [6]. In [6] the authors discuss the strengths and weaknesses of two perspectives of the web of data (the BTC dataset and of what they call the CKAN/LOD cloud metadata[v]) and propose high-quality collection of Linked Data snapshots for the community to gain a better insight into the underlying principles of dynamicity on the Web of Data.

Given that we find very few work on analysing the Web of Data from the perspective of multilinguality, in the next section we propose a methodology for performing our study.

3. Method and rationale

After analysing aforementioned corpora and performing several analysis, our initial candidate for extracting statistics and issuing questions will be the Dynamic Linked Data Observatory, being the most important reasons behind our decision the following: (1) it has reasonable size, (2) it is updated frequently updated so we can periodically run our analysis, (3) it tackles some of the issues found in BTC and LOD.

---

[i] The authors used the Billion Triple Challenge Dataset 2010 (see http://km.aifb. kit.edu/projects/btc-2010/)

[ii] http://stats.lod2.eu

[iii] http://stats.lod2.eu/languages

[iv] http://swse.deri.org/DyLDO/

[v] The CKAN(Comprehensive Knowledge Archive Network) repository contains a group lodcloud which is the one used in the creation of the LOD cloud.

After selecting the corpora, we have set up an infrastructure based on Apache Hadoop and Apache Pig that allows us to periodically analyse the data (i.e. every time a new corpora gets published) and run the different questions that we want to answer.

Having the dataset and the infrastructure, the method will be the following: (1) every time we want gather statistics on a new feature, we create a simple script and store it, (2) every time a new "observatory corpus" is published the stored scripts are executed, (3) the results can be analysed and published for the community.

We are currently in the first steps of this effort, but we are able to share some of our results and we would like to make this a community effort where researchers can suggest new studies and perspectives. During the seminar we would like to share some of our results, validate our current questions and gather new ones from other interested participants.

Acknowledgements

This work is supported by the Spanish Project TIN2010-17550 for the BabeLData project.

References

[1] J. Gracia, E. M. Ponsoda, P. Cimiano, A. G. Pérez, P. Buitelaar, and J. McCrae, "Challenges for the multilingual Web of Data," Journal of Web Semantics, vol. 11, pp. 63-71, Mar. 2012. Available at http://oa.upm.es/8848/1/Multiling.pdf

[2] L. Ding and T. Finin. Characterizing the semantic web on the web. In Proceedings of the 5th International Semantic Web Conference, 2006.

[3] M. d'Aquin, C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, and E. Motta. Characterizing knowledge on the semantic web with watson. In R. Garcia-Castro, D. Vrandecic, A. Gmez-Prez, Y. Sure, and Z. Huang, editors, EON, volume 329 of CEUR Workshop Proceedings, pages 1{10. CEUR-WS.org, 2007.

[4] Ell, B., Vrandecic, D., Simperl, E.P.B.: Labels in the web of data. In Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N.F., Blomqvist, E., eds.: International Semantic Web Conference (1). Volume 7031 of Lecture Notes in Computer Science., Springer (2011) 162–176

[5] L. Ding, J. Shinavier, Z. Shangguan, and D. McGuinness. SameAs networks and beyond: Analyzing deployment status and implications of owl:sameAs in linked data. In Proceedings of ISWC, pages 145–160, 2010.

[6] Tobias Käfer, Jürgen Umbrich, Aidan Hogan and Axel Polleres, Towards a Dynamic Linked Data Observatory, in the Proceedings of the Linked Data on the Web WWW2012 Workshop (LDOW 2012), Lyon, France, 16 April, 2012.