The possibilities and challenges of using linked data for academic research: the case of the Talk of Europe project

Laura Hollink
VU University Amsterdam /
Centrum Wiskunde & Informatica
l.hollink@cwi.nl

Martijn Kleppe
Erasmus University Rotterdam
kleppe@eshhc.eur.nl

Max Kemman
University of Luxembourg
max.kemman@uni.lu

Astrid van Aggelen
VU University Amsterdam
a.e.van.aggelen@vu.nl

Willem van Hage
SynerScope
willem.van.hage@synerscope.com

The Talk of Europe project has made the proceedings of the plenary meetings of the European Parliament available as Linked Open Data, a way of publishing and connecting data on the Web. Access to the records of what happens during the meetings of the European Parliament (EP) is a crucial part of democracy. In addition, the proceedings are valuable source material for scholars in history, politicology (Proksch, 2010), natural language processing (Nusselder, 2009) and machine translation (Koehn, 2005). However, the EP web portal only offers limited search functionality. By publishing this data as Linked Open Data, we aim to improve access for scholars. **In this presentation we will reflect on the benefits and implications of linking the proceedings of the EP to a number of other datasets. Additionally, during the demonstration session, we will show how the SynerScope visualization tool enables an exploration of the links within and across datasets.**

Up until now, we have linked the proceedings to four external datasets: [1] a database of professional affiliations of the members of the EP (Høyland, 2009), [2] DBpedia[1], the semantic web mirror of Wikipedia, [3] Geonames[2], a geographical thesaurus, and (4) the politicians and parties of the parliament of Italy[3]. Through these links, we enable scholars to access and use the knowledge that is captured in these external datasets. For example: the former member of parliament Jeanine Hennis is linked to her entry in Høyland's database telling us that she was a member of the Committee on

[1] http://dbpedia.org/
[2] http://www.geonames.org/
[3] http://data.camera.it/

Transport and Tourism; to her DBpedia page, giving access to her birthdate and place, diplomas, and jobs outside the EP; the country that she represents (The Netherlands) is linked to its corresponding Geonames entity, providing information on population density, income and neighbouring countries. These links enable queries that are not possible on either of these datasets alone. For example, if we combine birthplace information from DBpedia with Geonames' geographical information, we can query for members of the EP that were born outside Europe.

During two creative camps, we invited teams of scholars to use our linked dataset. Interesting applications were built on top of it, showing the possibilities of Linked Open Data for scholars. However, at the same time, an increase in the number of external datasets that is used raises questions about the correctness, transparency, stability and completeness of the results, which are fundamental questions for humanities researchers, for whom provenance is crucial. We furthermore observed that while the organisation of the creative camps stimulated the uptake of the dataset, the Linked Open Data format remains a hurdle for many humanities scholars.

References

Proksch, S.-O. and Slapin, J.B. (2010) Position taking in european parliament speeches. *British Journal of Political Science*, Vol. 40, Issue 03, pp. 587–611.

Nusselder, A., Peetz, H., Schuth, A. and Marx, M. (2009) Helping people to choose for whom to vote. A web information system for the 2009 european elections. *Proceedings of the 18th ACM conference on Information and knowledge management*, ACM, pages 2095–2096.

Koehn, P. (2005) Europarl: A parallel corpus for statistical machine translation. *MT summit*, volume 5, pages 79–86.

Høyland, B., Sircar, I. and Hix, S. (2009) An Automated Database of the European Parliament. *European Union Politics*, Vol 10, Issue 1, pp. 143-152 .