Abstract for the 16th European Networked Knowledge Organization Systems (NKOS) Workshop

Subject data in national bibliographies published as Linked data

Kim Tallerås

Department of Archivistics, Library and Information Science,
Oslo and Akershus University College

In recent years, vast amounts of library data have been transformed and published on the web according to Linked data principles. Much has been written about the benefits of transforming metadata according to such principles, and many issues have been thoroughly discussed, while less effort has been spent on evaluations of the published results. This study will examine how four European National libraries (Bibliothèque nationale de France, the British Library, Biblioteca Nacional de España and Deutsche Nationalbibliothek) have treated subject data in their transformation of national bibliographies to Linked data.

The study covers Linked data practices such as interlinking to external sources, modelling and (re)use of vocabularies. These practices will be evaluated against established statistical metrics found in the literature on Linked data quality (see Zaveri et al. (2015) for an overview, and Schmachtenberg, Bizer, & Paulheim (2014) for a concrete application). In addition, an investigation of potential heterogeneity conflicts (as described by Haslhofer & Klas (2010)) will be conducted.

Available dumps of the national bibliographies in RDF format have been downloaded and ingested into a local triple store. SPARQL are used to perform statistical analysis of the entire data sets, and to extract sample data which enable a "close reading" of RDF triples surrounding a limited selection of works found in all sets.

Initial findings from the corpus statistics shows that subject data constitutes a significant portion of the published data (on average almost 10% of all resources (with a class membership) are members of skos:Concept or a similar class). There are few direct links between the sets, but many indirect connections via common links to external sources, such as Dewey and LCSH.

The linked data sets have implemented FRBR entities such as Work and Expression in varying degree. To reduce the amount of potential heterogeneity conflicts across the sets and across the sets and external sources, there may be need to conduct a proper mapping of (FRBR) entity types in order to exploit connections and build proper applications on top of them.

The study is part of a PhD project on interoperability of linked library data, and relates to workshop themes such as "KOS Linked data" and "Evaluation of KOS-based systems".

References:

Haslhofer, B., & Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, *42*(2), 1–37. http://doi.org/10.1145/1667062.1667064

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the Linked Data Best Practices in Different Topical Domains. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, … C. Goble (Eds.), *ISWC 2014, LNCS 8796* (pp. 245–260). Cham: Springer http://doi.org/10.1007/978-3-319-11964-9_16

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2015). Quality assessment for Linked Data: A Survey. *Semantic Web*, *7*(1), 63–93. http://doi.org/10.3233/SW-150175