Interlinking Open Data on the Web

Chris Bizer[1], Tom Heath[2], Danny Ayers[3], and Yves Raimond[4]

[1] Freie Universität Berlin, `chris[at]bizer.de`
[2] Knowledge Media Institute, The Open University, `t.heath[at]open.ac.uk`
[3] Free Author, `danny.ayers[at]gmail.com`
[4] Centre for Digital Music, Queen Mary, University of
London, `yves.raimond[at]elec.qmul.ac.uk`

Abstract. A fundamental prerequisite of the Semantic Web is the existence of large amounts of meaningfully interlinked RDF data on the Web. The W3C SWEO community project *Linking Open Data* has made various open datasets available on the Web as RDF, and developed automated mechanisms to interlink them with RDF statements. Collectively, the datasets currently consist of over one billion triples. We believe that large scale interlinking will demonstrate the value of the Semantic Web compared to more centralized approaches such as Google Base[5]. This paper outlines the work to date and describes the accompanying demonstration.

A functioning Semantic Web is predicated on the availability of large amounts of data as RDF; not in isolated islands but as a Web of interlinked datasets. To date this prerequisite has not been widely met, leading to criticism of the broader endeavour and hindering the progress of developers wishing to build Semantic Web applications. Thanks to the Open Data movement, a valuable opportunity exists to partially rectify this situation by making existing royalty-free datasets (such as Wikipedia, Musicbrainz, Geonames, Wordnet, and DBLP) available as RDF, and interlinking them on a large scale.

The W3C SWEO[6] community project *Linking Open Data*[7] is pursuing this avenue, and has published several large interlinked RDF datasets on the Web. The project follows the Linked Data principles[8] outlined by Tim Berners-Lee, such that: all items of interest should be identified using URI references; all URI references should be resolvable on the Web to RDF descriptions; and every RDF triple is conceived as a hyperlink that can be followed by Semantic Web browsers and crawlers.

Our Web of interlinked datasets currently consists of dbpedia (91 million triples), Geonames (60 million triples), Musicbrainz (50 million triples), the dbtune music server (4 million triples), the DBLP bibliography (15 million triples),

[5] http://www.google.com/base
[6] http://www.w3.org/2001/sw/sweo/
[7] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[8] http://www.w3.org/DesignIssues/LinkedData.html

Revyu reviews and ratings (15 thousand triples), a US census dataset (700 million triples), and the RDF Book Mashup (several billion triples).

These datasets are interlinked by approximately 150.000 RDF links, in the form of triples that connect a subject URI from one dataset with an object URI from another dataset. Using these links one can navigate from a computer scientist in dbpedia to her publications in the DBLP database, from a dbpedia book to reviews and sales offers for this book provided by the RDF Book Mashup, or from a band in dbpedia to a list of their songs provided by Musicbrainz or dbtune.

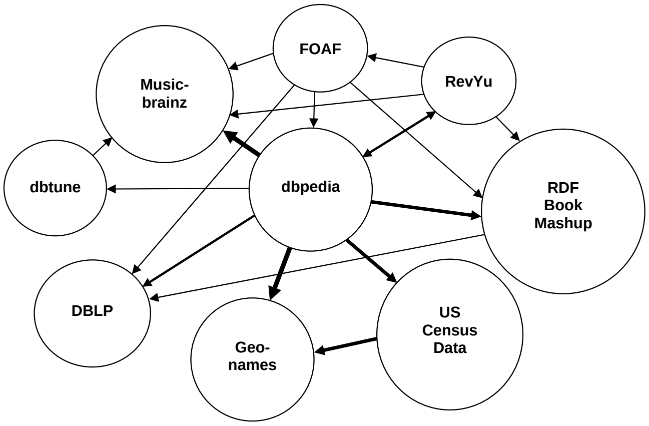

Fig. 1. Linking relationships within our web of datasets.

In our demonstration, we will show how this web of datasets is browsed using three different Semantic Web browsers: the Tabulator browser developed at MIT, the Disco browser developed at Freie Universität Berlin and the OpenLink Data Web browser. We will also demonstrate the Zitgist Semantic Web search engine which crawls the data and provides an integrated view on it as well as a easy-to-use search interface.

RDF is the obvious technology to interlink data from various data sources. The RDF datasets created by the project can be used as a testbed for Semantic Web browsers and crawlers, RDF stores and reasoning engines, as well as for data linkage, data cleansing, and data mining tools. We encourage people to set RDF links into our datasets, as each new link helps to bootstrap the Semantic Web as a whole.

In addition to the authors, the following people currently contribute to the project: Sören Auer, Josh Tauberer (University of Pennsylvania), Frederick Giasson (Zitgist), Kingsley Idehen, Orri Erling (OpenLink), Georgi Kobilarov, Richard Cyganiak (Freie Universität Berlin), Stefano Mazzocchi (MIT), Bernard Vatant (Mondeca), Marc Wick (Geonames). The *Linking Open Data* project is a community effort and we highly welcome further participants.