The DBpedia Data Stack

Towards a sustainable DBpedia Project to provide

a public data infrastructure for Europe

Dimitris Kontokostas and Sebastian Hellmann

Submitted to EDF as an **Impact contribution**

DBpedia is currently a very successful data dissemination project with high-industry uptake. We analysed the project and identified **several barriers preventing DBpedia from realizing its full potential** and ensuring a sustainable operation, namely:

1. lack of tools to support improved and cost-efficient data curation and multilingualism,
2. lack of highly available value-added services with quality of service (QoS) guarantees and lack of enterprise-optimized infrastructures
3. lack of proper documentation, tutorials and support, resulting in steep learning curves for new technologies

These obstacles prohibit the participation of SMEs in Linked Data environments, thus depriving them of valuable resources for business diversification and development. On the other hand, Linked Data technologies are stuck in their original research roots, being also deprived from real world development opportunities.

To address these barriers, **technological advances as well as an organizational framework are required** to provide a sustainable environment for future developments. Although the software underlying DBpedia received feature additions through several research project deliverables, the project itself remains unfunded and does not have a proper organisational structure. This is a commonplace among linked data sets, especially those originating from research projects. Opening the infrastructure to the community has been required to prevent stagnation and breakdown of data maintenance.

A step towards this effort is the first public DBpedia community meeting[1] that will take place in Amsterdam on January 2014. There we will try to identify stakeholders and their interests and look for ways to generate income to improve DBpedia further.

DBpedia Internationalization

The early versions of DBpedia used only the English Wikipedia as sole source. Since the beginning, the focus of DBpedia has been to build a fused, integrated data stack by integrating information from many different Wikipedia editions. The emphasis of this fused DBpedia was still on the English Wikipedia as it is the most abundant language edition. During the fusion process, however, language-specific information was lost or ignored. Recent internationalization development established best practices (complemented by software) that allow the DBpedia community to easily generate, maintain and properly interlink a language-specific DBpedia

[1] http://dbpedia.org/meetings

edition.

At the time of writing, 14 official DBpedia chapters, apart from the English one exist, namely: Basque, Czech, Dutch, French, German, Greek, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian and Spanish.[2] All these DBpedia language editions provide the same data stack as the English DBpedia. Additionally, as the English DBpedia serves as a central hub at the Web of Data, local DBpedia editions can may serve as a local hub for their respective country-level LOD clouds.

DBpedia Data Stack

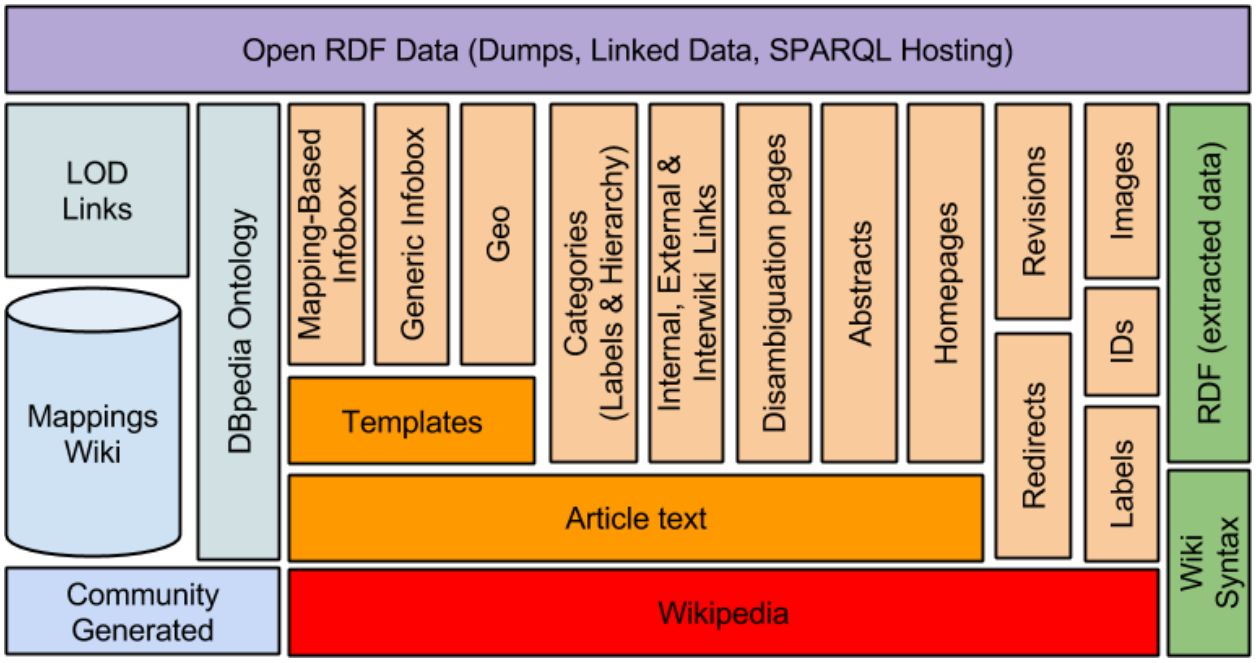

Figure 1: The DBpedia Data Stack

The DBpedia Data Stack is depicted in Figure 1. By taking a Wikipedia language edition as input, DBpedia can easily identify basic page information such as the page title (label), the page id, the revision number, redirects and page type. By parsing the Wiki Syntax of an article we get access to the Wikipedia categories and hierarchy, links (internal, external or to other wikis), article disambiguations, homepages and short and long article excerpts. One of the richest source of semi-structured information in Wikipedia articles tapped by DBpedia are templates. From templates we get geographical information (coordinates) as well as extract infobox information. For infoboxes, DBpedia offers a generic infobox extraction where data are generated greedily and are not normalized to strict datatypes. In the Mapping-Based infobox extraction, data normalization is crowdsourced through the mappings wiki[3] and mapped to the DBpedia ontology, thus provides very high quality data. Finally, links to the LOD cloud are also post-processed and offered in the data stack. All this information is available as RDF and accessible through Turtle and NTriples dumps[4], a SPARQL Endpoint[5] a Linked Data interface[6].

[2] http://dbpedia.org/Internationalization
[3] http://mappings.dbbpedia.org
[4] http://downloads.dbpedia.org
[5] http://dbpedia.org/sparql

Planned Data Stack extensions

Natural Language Processing

Recently, the DBpedia community has experienced an immense increase in activity and we believe, that the time has come to explore the connection between DBpedia & Natural Language Processing (NLP) in a yet unprecedented depth. DBpedia has a long-standing tradition to provide useful data as well as a commitment to reliable Semantic Web technologies and living best practices. As the extraction of Wikipedia's infoboxes by DBpedia matures, we can shift our focus on new challenges such as extracting information from the unstructured article text as well as becoming a testing ground for multilingual NLP methods.

Abbreviation Base

Abbreviation lists can be obtained by using the Web of Data and the open and multilingual resources of DBpedia and the DBpedia Wiktionary extraction. There is a number of advantages this approach has in favour of precompiled lists or using classical dictionaries as only data source:

- DBpedia as a resource derived from Wikipedia is a free and open data source that leverages the crowd as a means to ensure comprehensiveness and quality of the data
- Using those resources provides a solid base of relevant (as per Wikipedia standards) data for free
- The approach is inherently multilingual, as the DBpedia exists for 170 languages
- Additional information besides the abbreviation string is provided and can be used for disambiguation

In DBpedia, like in Wikipedia, the redirect links serve to link alternative names of resources to a main resource or to a disambiguation resource, that links different meanings of the string in question. Abbreviations belong, in most cases, to one of these groups.

Lexical Resources and Terminologies for Question Answering

As the body of knowledge available as linked data grows, so does the need to provide methods that make this knowledge accessible for humans. Such methods usually require knowledge about how the vocabulary elements used in the available ontologies and datasets are verbalized

in natural language. This has lead to much interest in the development of models and frameworks for publishing ontology lexica as linked data. In this paper we describe a process for the manual development of such lexica in lemon format and illustrate some of the key challenges involved. As a proof of concept, Unger et al. 2013[7] provide a manually created English lexicon for the DBpedia ontology and describe its first release.

[6] e.g. http://dbpedia.org/resource/Europe
[7] http://ceur-ws.org/Vol-1064/Unger_lemon.pdf

Further Data

In its version 3.9 DBpedia already included links to **WikiData** into its Data Stack[8]. At the time of writing even more experimental data from WikiData is loaded into the main endpoint. Furthermore the integration of **Wikimedia Commons**[9] is planned. The extraction of data from **Wiktionary** has already taken place and is integrated into the data dissemination structure of DBpedia[10]

Quality Assessment

Data quality is a top priority for DBpedia. We are already working on a Test-Driven data quality assessment approach[11] where we define automated and manual data quality test cases and run them regularly against our data stack. This will ensure a basic level of quality in DBpedia and ensure previously addressed quality issues will be quickly identified if they reappear.

CVs

Dimitris Kontokostas (AKSW, Universität Leipzig, Germany, kontokostas@informatik.uni-leipzig.de, http://aksw.org/DimitrisKontokostas) is a researcher at the AKSW research group. He finished his Master thesis in 2012 at University of Thessaloniki and his Bachelor thesis in 2004 at Technical University of Crete. Between his bachelor and Master he worked as a programmer and an informatics teacher. Dimitris is now an active maintainer of the DBpedia project and a founding member of the DBpedia Internationalization Effort. His research focuses on improving the quality of DBpedia and Linked Open Data in general.

Sebastian Hellmann (AKSW, Universität Leipzig, Germany, hellmann@informatik.uni-leipzig.de, http://bis.informatik.uni-leipzig.de/SebastianHellmann) finished his Master thesis in 2008 at University of Leipzig and is currently a research fellow for the AKSW research group and also a member of the LOD2 EU project. He is contributor, co-founder and leader of several open source projects including DL-Learner, DBpedia, and NLP2RDF. Sebastian is author of over 15 peer-reviewed scientific publications and was chair at the Open Knowledge Conference in 2011, the Workshop on Linked Data in Linguistics 2012, the Linked Data Cup 2012 and the Multilingual Linked Data for Enterprises 2012 workshop.

---

8 http://blog.dbpedia.org/2013/09/17/dbpedia-39-released-including-wider-infobox-coverage-additional-type-statements-and-new-yago-and-wikidata-links/

9 http://commons.wikimedia.org

10 http://dbpedia.org/Wiktionary

11 http://svn.aksw.org/papers/2014/WWW_Databugger/public.pdf