Where there's a Wiki, there's a way: DBpedia's linked data solution to an encylopaedia's problem

Class: Digital Preservation
Lecturer: Kara Van Malssen
Student: David Neary

The pivotal role that Wikipedia has come to play in the Information Age cannot be denied, however much it may be maligned by some. In 2013 Wikipedia was reported as being the sixth most accessed website in the world, as well as the most widely used encyclopaedia.[1] While constant vigilance on behalf of editors to swiftly correct errors and deflate opinions-as-facts is the most obvious way to maintain Wikipedia as a serious and respectable source of information, others in the information sciences have sought to increase and improve what Wikipedia can do and offer.

DBpedia is a "crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web"[2] in order to enable more complex searches to be achieved. In simpler terms, DBpedia mines Wikipedia for machine-readable facts, and allows those facts to be collated with and against others in order to produce new ways of searching Wikipedia. The DBpedia ('Database-pedia') Project was begun by researchers at the University of Leipzig and the Free University of Berlin in the mid-2000s, in collaboration with OpenLink Software. It was launched in 2007 under the same licenses that Wikipedia is run under, the Creative Commons Attribution-ShareAlike 3.0 License.

The concept under which DBpedia functions is deceptively simple. DBpedia extracts structured information from Wikipedia articles, creating DBpedia entities with listed properties (fact categories) and values (the associated facts). More specifically, DBpedia uses RDF (Resource Description Framework), "the main building block of the Semantic Web",[3] as the specification in which to represent the extracted data. Using a conceptual framework known as triples which link subjects to categorical roles (e.g. actor, song, country) or other defining linked information, RDF connects all subjects to its values and their properties, making it particularly suitable for DBpedia's data-mining project.

---

[1] Lehmann, Isele, Jakob, Jentzsch, Kontokostas, Mendes, Hellman, Morsey, van Kleef, Auer, Bizer, 'DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia', 2013. http://svn.aksw.org/papers/2013/SWJ_DBpedia/public.pdf p. 2

[2] http://wiki.dbpedia.org/About

[3] Morsey, Lehmann, Auer, Stadler, Hellman, 'DBpedia and the Live Extraction of Structured Data from Wikipedia', Leipzig, 2012. http://jens-lehmann.org/files/2012/program_el_dbpedia_live.pdf

Much of this information is taken from the individual article's infobox, the information table located in the top right corner of many Wikipedia pages. Many 'properties' and 'values' are clearly listed here, which DBpedia reads and comprehends – this allows that information to be searched in such a way that the subject of the article can be easily turned up.

Thus, for example, one could search for a movie, produced in Spain between 1990 and 1999 and DBpedia would list every film that Wikipedia registers as having been produced in Spain during those dates. One could even add a minimum running time of 90 minutes to exclude shorter films. Data links are thus created through DBpedia's reading of the information in the infoboxes of all of these films. Linked data allows the search to travel backward, as it were, to every film that lists 'Spain' (value) as its country (property), and similarly searching the years and running times. This is what is known as a 'faceted Wikipedia search'. These "allow users to ask complex questions [of Wikipedia]… In order to answer such questions, a search engine must facilitate structured knowledge which needs to be extracted from the underlying articles."[4]

The usefulness of this restructuring of information cannot be underestimated, as it allows information to be restructured based on the query, a task that a Google search is incapable of performing on its own. For example, using DBpedia with map software (such as Google Maps), one can instantly create a map of Europe with the capital cities of every country pinpointed, as seen in the **Fig.1** below. DBpedia would be able to access the countries that make up Europe, and harvest their capitals from their infoboxes, then use the coordinates listed on each capital's page to name and pinpoint the markers on the map.

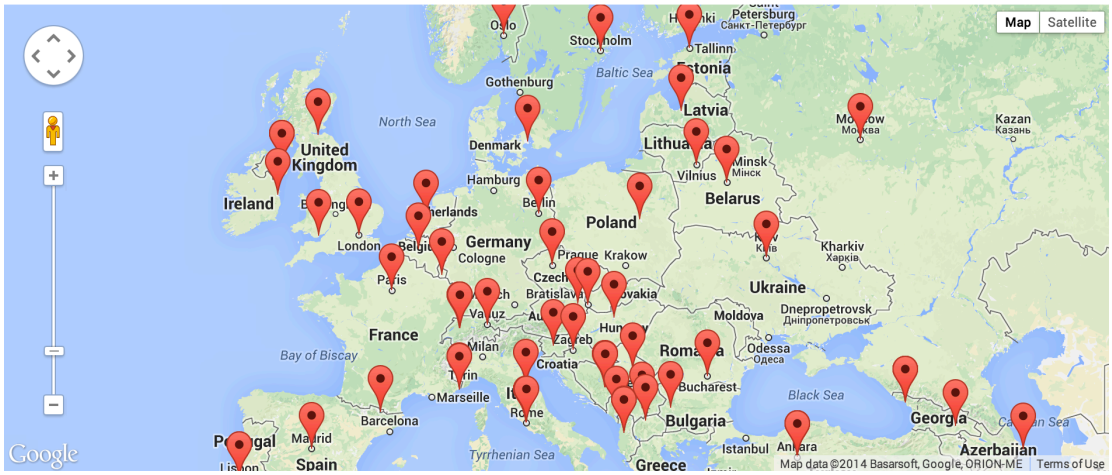

Fig.1 (via http://liris.cnrs.fr/~pchampin/spark/gmapv3.html)

DBpedia can be used to create datasets and tables all using information taken from Wikipedia articles. Even more impressively, as of Version 2014, DBpedia can now fully access Wikipedia in 125 different languages, using the same linked datasets. This is useful for users of Wikipedia in smaller editions, as information will be taken from the English and other large editions to bolster the

[4] Hahn, Bizer, Sahnwaldt, Herta, Robinson, Bürgle, Düwiger, Scheel, 'Faceted Wikipedia Search'. http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/hahn-etal-faceted-wikipedia-search-BIS2010.pdf p. 1

DBpedia file. Similarly, "the larger DBpedia editions can benefit from more specialized knowledge from localized editions, such as data about smaller towns which is often only present in the corresponding language editions".[5]

Another recent improvement is the game-changing introduction of DBpedia-Live. The "Semantic Web mirror of Wikipedia"[6], DBpedia-Live means information in DBpedia is updated within minutes of changes made to the original Wikipedia article. According to Lehmann et al, "Wikipedia articles are continuously revised at a very high rate, e.g. the English Wikipedia, in June 2013, has approximately 3.3 million edits per month which is equal to 77 edits per minute."[7] Before DBpedia-Live, DBpedia could languish behind Wikipedia for weeks at a time, but new tools utilising the Open Archives Initiative Protocol for Metadata Harvesting allow it to keep updated almost simultaneously.

While DBpedia can be used in conjunction with a variety of softwares, all of the information it has mined from Wikipedia can be found by using the address dbpedia.org/page/[Wikipedia page], where [Wikipedia page] is standing in for the subject entry on Wikipedia exactly as it is written in the address bar.

The future plans for DBpedia seem achievable. The project's plan to provide a "proof-of-concept" for future potential large-scale data extraction projects[8] seems more or less complete. More complicated but crucial is to develop their software so that it can extract semantic information from typed text; this would make Wikipedia articles considerably deeper mines of data for DBpedia. It is also hoped that using DBpedia-Live and reading automatically linked data that DBpedia might also serve as an automated fact-checker and provider of error feedback for Wikipedia. With DBpedia reliant on the information from Wikipedia, it would be remarkable for it to give back to Wikipedia in this way, making the relationship between the two a symbiotic one.

---

[5] Lehmann et al, op. cit., p. 6
[6] http://wiki.dbpedia.org/DBpediaLive
[7] Lehmann et al, op. cit., p. 14
[8] Lehmann et al, p. 25

Webography

Auer, Bizer, Lehmann, Kobilarov, Cyganiak, Ives, 'DBpedia: A Nucleus for a Web of Open Data' in *The Semantic Web*, ed. Aberer et al., Busan, 2007. http://www.informatik.uni-leipzig.de/~auer/publication/dbpedia.pdf

Becker, 'DBpedia – Extracting structured data from Wikipedia', Buenos Aires, 2009. http://wikimania2009.wikimedia.org/wiki/Proceedings:174

Bizer, Lehmann, Kobilarov, Auer, Becker, Cyganiak, Hellmann, 'DBpedia – A Crystallization Point for the Web of Data'. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, Issue 7, 2009. http://jens-lehmann.org/files/2009/dbpedia_jws.pdf

Daiber, Jakob, Hokamp, Mendes, 'Improving efficiency and accuracy in multilingual entity extraction' in *Proceedings of the 9th International Conference on Semantic Systems,* 2013.

Hahn, Bizer, Sahnwaldt, Herta, Robinson, Bürgle, Düwiger, Scheel, 'Faceted Wikipedia Search', Berlin, 2010. http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/hahn-etal-faceted-wikipedia-search-BIS2010.pdf

Jentzsch, 'DBpedia – Extracting structured data from Wikipedia', Cologne, 2009. http://www.anjajentzsch.de/slides/SWIB09_DBpedia.pdf

Lehmann, Isele, Jakob, Jentzsch, Kontokostas, Mendes, Hellman, Morsey, van Kleef, Auer, Bizer, 'DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia', 2013. http://svn.aksw.org/papers/2013/SWJ_DBpedia/public.pdf

Mendes, Jakob, Bizer. 'DBpedia for NLP: A Multilingual Cross-domain Knowledge Base', Istanbul.

Mendes, Jakob, García-Silva, Bizer. 'DBpedia Spotlight: Shedding Light on the Web of Documents', Graz, 2011.

Morsey, Lehmann, Auer, Stadler, Hellman, 'DBpedia and the Live Extraction of Structured Data from Wikipedia', Leipzig, 2012. http://jens-lehmann.org/files/2012/program_el_dbpedia_live.pdf