Towards a Korean DBpedia and an Approach for Complementing the Korean Wikipedia based on DBpedia

Eun-kyung Kim[1], Matthias Weidl[2], Key-Sun Choi[1], Sören Auer[2]

[1] Semantic Web Research Center, CS Department, KAIST, Korea, 305-701
[2] Universität Leipzig, Department of Computer Science, Johannisgasse 26, D-04103
Leipzig, Germany
kekeeo@world.kaist.ac.kr, kschoi@world.kaist.ac.kr
mam07jct@studserv.uni-leipzig.de, auer@informatik.uni-leipzig.de

Abstract. In the first part of this paper we report about experiences when applying the DBpedia extraction framework to the Korean Wikipedia. We improved the extraction of non-Latin characters and extended the framework with pluggable internationalization components in order to facilitate the extraction of localized information. With these improvements we almost doubled the amount of extracted triples. We also will present the results of the extraction for Korean. In the second part, we present a conceptual study aimed at understanding the impact of international resource synchronization in DBpedia. In the absence of any information synchronization, each country would construct its own datasets and manage it from its users. Moreover the cooperation across the various countries is adversely affected.

Keywords: Synchronization, Wikipedia, DBpedia, Multi-lingual

1 Introduction

Wikipedia is the largest encyclopedia of mankind and is written collaboratively by people all around the world. Everybody can access this knowledge as well as add and edit articles. Right now Wikipedia is available in 260 languages and the quality of the articles reached a high level [1]. However, Wikipedia only offers full-text search for this textual information. For that reason, different projects have been started to convert this information into structured knowledge, which can be used by Semantic Web technologies to ask sophisticated queries against Wikipedia. One of these projects is DBpedia [2], which stores structured information in RDF. DBpedia reached a high-quality of the extracted information and offers datasets in 91 different languages. However, DBpedia lacks sufficient support for non-English languages. For example, DBpedia only extracts data from non-English articles that have an interlanguage link[3], to an English article.

[3] http://en.wikipedia.org/wiki/Help:Interlanguage_links

Therefore, data which could be obtained from other articles is not included and hence cannot be queried. Another problem is the support for non-Latin characters which among other things results in problems during the extraction process. Wikipedia language editions with a relatively small number of articles (compared to the English version) could benefit from an automatic translation and complementation based on DBpedia. The Korean Wikipedia, for example, was founded in October 2002 and reached ten thousand articles in June 2005[4] . Since February 2010, it has over 130,000 articles and is the 21st largest Wikipedia[5]. Despite of this growth, compared to the English version with 3.2 million articles it is still small.

The goal of this paper is two-fold: (1) to improve the DBpedia extraction from non-Latin language editions and (2) to automatically translate information from the English DBpedia in order to complement the Korean Wikipedia.

The first aim is to improve the quality of the extraction in particular for the Korean language and to make it easier for other users to add support for their native languages. For this reason the DBpedia framework will be extended with a plug-in system. To query the Korean DBpedia dataset a Virtuoso Server[3] with a SPARQL endpoint will be installed. The second aim is to translate infoboxes from the English Wikipedia into Korean and insert it into the Korean Wikipedia and consequently the Korean DBpedia as well. In recent years, there has been significant research in the area of coordinated control of multi-languages. Although English has been accepted as a global standard to exchange information between different countries, companies and people, the majority of users are attracted by projects and web sites if the information is available in their native language as well. Another important fact is that various Wikipedia editions from different languages can offer more precise information related to a large number of native speakers of the language, such as countries, cities, people and culture. For example, information about 김재규 (transcribing into Kim Jaegyu), a music conductor of the Republic of Korea, is only available in Korean and Chinese at the moment.

The paper is structured as follows: In Section 2, we give an overview about the work on the Korean DBpedia. Section 3 explains the complementation of Korean Wikipedia using DBpedia. In Section 4, we review related work. Finally, we discuss future work and conclude in Section 5.

2 Building the Korean DBpedia

The Korean DBpedia uses the same framework to extract datasets from Wikipedia as the English version. However, the framework does not have sufficient support for non-English languages, especially for the non-Latin alphabet based languages.

For testing and development purposes, a dump of the Korean Wikipedia was loaded into a local MySQL database. The first step was to use the current DBpedia extraction framework in order to obtain RDF triples from the database.

[4] http://stats.wikimedia.org/EN/ChartsWikipediaKO.htm

[5] http://meta.wikimedia.org/wiki/List_of_Wikipedias

At the beginning the focus was on infoboxes, because infobox templates offer already semi-structured information. But instead of just extracting articles that have a corresponding article in the English Wikipedia, like the datasets provided by DBpedia, all articles have been processed. More information about the DBpedia framework and the extraction process can be found in [4] and [5].

After the extraction process and the evaluation of the RDF triples, encoding problems have been discovered and fixed. In fact most of these problems will occur not only in Korean, but for all languages with non-Latin characters. Wikipedia and DBpedia use the UTF-8 and URL encoding. URI's in DBpedia have the form **http://ko.dbpedia.org/resource/Name**, where Name is taken from the URL of the source Wikipedia article, which has the form **http://ko.wikipedia.org/wiki/Name**. This approach has certain advantages. Further information can be found in [5]. For example an URI for 괴팅겐 (transcribing into Göttingen), as a property in a RDF Triple, would look as follows:

http://ko.dbpedia.org/property/%EA%B4%B4%ED%8C%85%EA%B2%90

This property URI contains the "%" character and thus cannot be serialized as RDF/XML. For this reason another way has to be found to represent properties with "%" encoding. The solution in use by DBpedia is to replace "%" with "_percent_". This resulted in very long and confusing properties which also produced errors during the extraction process. This has not been a big issue for the English DBpedia, since it contains very few of those characters. For other languages this solution is unsuitable. To solve it, different solutions have been discussed. The first possibility is to just drop the triples that contain such characters. Of course this is not an applicable solution for languages that mainly consist of characters which have to be encoded. The second solution was to use a shorter encoding but with this approach the Wikipedia encoding cannot be maintained. Another possibility is to use the "%" character and add an underscore at the end of the string. With this modification, the Wikipedia encoding could be maintained and the RDF/XML can be serialized. At the moment we use this solution during the extraction process. The use of IRI[6]'s instead of URI's is another possibility which we will discuss in Section 5. An option has been added to the framework configuration to control which kind of encoding should be used.

Because languages differ in grammar as well, it is obvious that every language uses its own kind of formats and templates. For example, dates in the English and in the Korean Wikipedia look as follows:

English date format: 25 October 2009
Korean date format: 2009년 10월 25일

For that reason every language has to define its own extraction methods. To realize this, a plug-in system has been added to the DBpedia extraction framework (see Fig. 1).

[6] http://www.w3.org/International/articles/idn-and-iri/

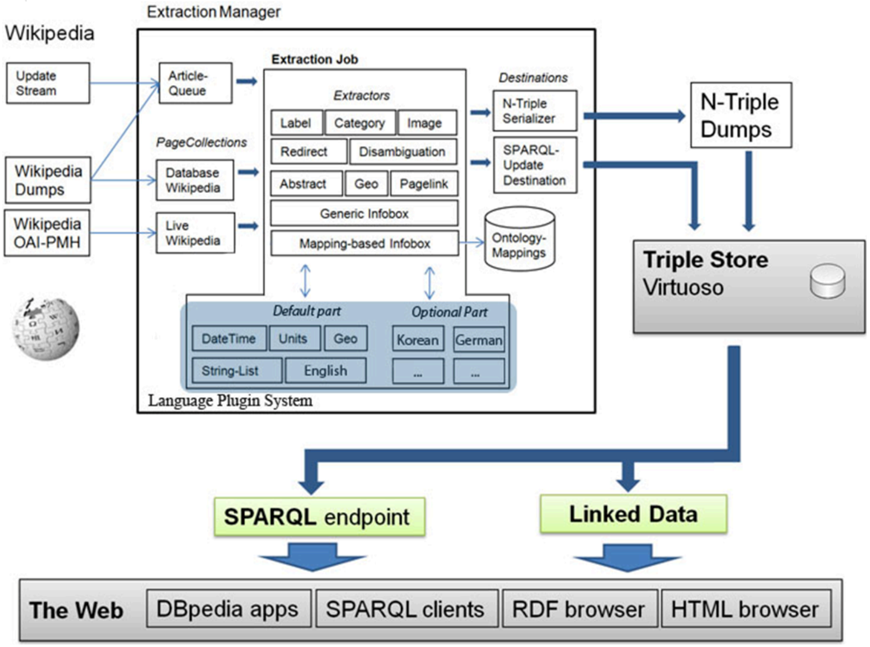


Fig. 1. DBpedia framework with the language plug-in system

The plug-in system consists of two parts: the default part and an optional part. The default part contains extraction methods for the English Wikipedia and functions for datatype recognition, for example, currencies and measurements. This part will always be applied first, independent from which language is actually extracted.

The second part is optional. It will be used automatically if the current language is not English. The extractor will load the plug-in file for the corresponding language if it exists. If the extractor did not find a match in the default part, it will use the optional part to check the current string for corresponding templates. The same approach is used for sub-templates which are contained in the current template.

After these problems have been resolved and the plug-in system has been added to the DBpedia extraction framework, the dataset derived from the Korean Wikipedia infoboxes consists of more than 103,000 resource descriptions with more than 937,000 RDF triples in total. The old framework only extracted 55,105 resource descriptions with around 485,000 RDF triples. The amount of triples and templates was almost doubled. The extended framework also extracted templates which have not been extracted by the old framework at all. A comparison between some example templates extracted by the old framework and the extended version can be found in Fig. 2.

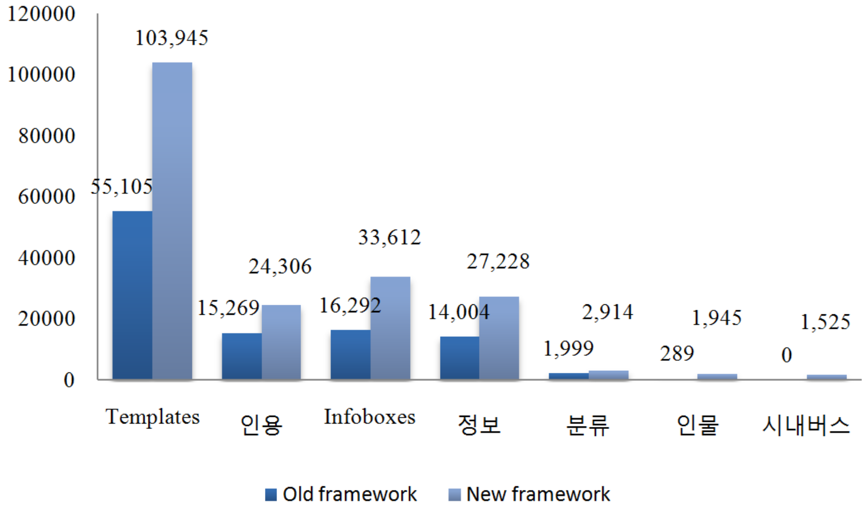


Fig. 2. Comparison between the old and the extended framework

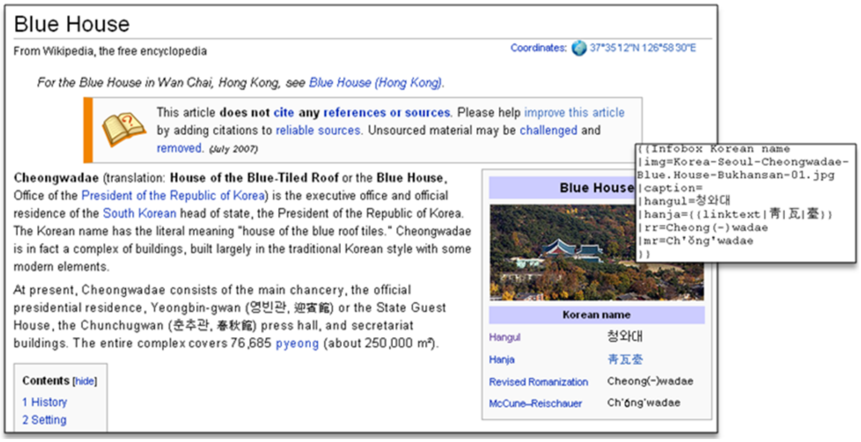
Table 1. The Korean DBpedia dataset

Extractor	**Description**	**Triples**
Abstract	Extracts the abstract of an article.	362K
ArticleCategories	Extracts the Wikipedia categories an article belongs to.	277.8K
Categories	Information about which concept is a category and how categories are related to each other.	40.9k
Disambiguation	Extracts the disambiguation links from a Wikipedia page.	79.8K
Externallinks	Extracts all links from the “External Links” section of a Wikipedia article.	105.4K
Geocoordinates	Extracts geo information of articles.	3.6K
Image	Extracts the first image of a Wikipedia page with a thumbnail and the full size image.	91.4K
Infobox	Extracts all information from Wikipedia infoboxes.	1.106K
Label	Extracts the pagelabel from a Wikipedia page.	182.6K
Pagelinks	Extracts all internal links of an article.	2.64M
Redirects	Extracts redirects in Wikipedia articles to identify synonymous terms.	40.2k
SKOS	This extractor represents Wikipedia categories using the SKOS vocabulary.	81.9k
Wikipage	For every DBpedia resource, this extractor sets a Link to the corresponding Wikipedia page.	182.6K
Total		5.21M

We already started to extend the support for other extractors. Until now the extractors mentioned in Table 1 are supported for Korean.

3 Complementation of the Korean Wikipedia using DBpedia

The infobox is manually created by authors that create or edit an article. As a result, many articles have no infoboxes and other articles contain infoboxes which are not complete. Moreover, even the interlanguage linked articles do not use the same infobox template or contain different amount of information. This is what we call the imbalance of information. This problem raises an important issue about multi-lingual access on the Web. In the Korean Wikipedia multi-lingual access is prevented due to the lack of interlanguage links.



Blue House

From Wikipedia, the free encyclopedia

Coordinates: 37°35′12″N 126°58′30″E

For the Blue House in Wan Chai, Hong Kong, see Blue House (Hong Kong).

This article **does not cite any references or sources**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(July 2007)*

Cheongwadae (translation: **House of the Blue-Tiled Roof** or the **Blue House**, Office of the President of the Republic of Korea) is the executive office and official residence of the South Korean head of state, the President of the Republic of Korea. The Korean name has the literal meaning "house of the blue roof tiles." Cheongwadae is in fact a complex of buildings, built largely in the traditional Korean style with some modern elements.

At present, Cheongwadae consists of the main chancery, the official presidential residence, Yeongbin-gwan (영빈관, 迎賓館) or the State Guest House, the Chunchugwan (춘추관, 春秋館) press hall, and secretariat buildings. The entire complex covers 76,685 pyeong (about 250,000 m²).

Contents [hide]
1 History
2 Setting

Blue House

Korean name	
Hangul	청와대
Hanja	靑瓦臺
Revised Romanization	Cheong(-)wadae
McCune–Reischauer	Ch'ŏng'wadae

```
{{Infobox Korean name
|img=Korea-Seoul-Cheongwadae-
Blue.House-Bukhansan-01.jpg
|caption=
|hangul=청와대
|hanja={{linktext|靑|瓦|臺}}
|rr=Cheong(-)wadae
|mr=Ch'ŏng'wadae
}}
```

Fig. 3. Wikipedia article and infobox about "Blue House" in English

For example (see Fig. 3), the "Blue House"[7] article in English contains the infobox template *Korean_Name*. However, the "Blue House" could be regarded as either a *Building* or a *Structure*. The "White House" article, which is very similar to the former, uses the *Historic Building* infobox template. Furthermore, the interlanguage linked "청와대" ("Blue House" in Korean) page does not contain any infobox.

There are several types of imbalances of infobox information between two different languages. According to the presence of infobox, we classified interlanguages linked pairs of articles into three groups: Short Infobox (S), Distant Infobox (D), and Missing Infobox (M):

[7] The "Blue House" is the executive office and official residence of the South Korean head of state, the President of the Republic of Korea.

- The ***S-group*** contains pairs of articles which use the same infobox template but have a different amount of information. For example, an English-written article and a non-English-written article, which have an interlanguage link and use the same infobox template, but have a different amount of template attributes.
- The ***D-group*** contains pairs of articles which use different infobox templates. D-group emerges due to the different degrees of each Wikipedia communities' activities. In communities where many editors lively participate, template categories and formats are more well-defined and more fine-grained. For example, *philosopher*, *politician*, *officeholder* and *military person* templates in English are matched just *person* template in Korean. It appears not only a Korean Wikipedia but also non-English Wikipedias.
- The ***M-group*** contains pairs of articles where an infobox exists on only one side.

As a first step, we concentrate on S-group and D-group. We tried to enrich the infobox using dictionary-based term translation. In this work, DBpedia's English triples in infoboxes are translated into Korean triples. We used bilingual dictionary which is originally created for English-to-Korean translations from Wikipedia interlanguage links. Then we added translation patterns for multi-terms using bilingual alignment of collocations. Multi-terms are set of single terms such as *Computer Science*.

DBpedia[2] is a community which harvests the information from infoboxes. We translated English DBpedia into Korean. We also developed the Korean infobox extraction module in Python. This module identifies records contained in infoboxes, and then parse out the needed fields. A comparison of datasets is as follows:

- English Triples in DBpedia: 43,974,018
- Korean Dataset (Existing Triples/Translated Triples): 354,867/12,915,169

We can get translated Korean triples over 30 times larger than existing Korean triples. However, a large amount of translated triples has no predefined templates in Korean. There may be a need to form a template schema to organize the fine-grained template structure.

Thus we have built a template ontology, OntoCloud[8], from DBpedia and Wikipedia, which was released on September 2009, to efficiently build the template structure. The construction of OntoClolud consists of the following steps: (1) extracting templates of DBpedia as concepts in an ontology, for example, the Template:Infobox **Person** (2) extracting attributes of these templates, for example, *name* of Person. These attributes are mapped to properties in ontology. (3) constructing the concept hierarchy by set inclusion of attributes, for example, *Book* is a subclass of *Book series.*

- **Book_series** = {name, title_orig, translator, image, image_caption, author, illustrator, cover_artist, country, language, genre, publisher, media_type, pub_date, english_pub_date, preceded_by, followed_by}.

[8] http://swrc.kaist.ac.kr/ontocloud

- **Book** = {name, title_orig, translator, image, image_caption, author, illustrator, cover_artist, country, language, genre, publisher, pub_date, english_pub_date, media_type, pages, isbn, oclc, dewey, congress, preceded_by, followed_by}.

For the ontology building, similar types of templates are mapped to a concept. For example, the *Template:infobox_baseball_player* and *Template:infobox_asian_baseball_player* describe baseball player. Moreover different format of properties with same meaning should be refined, for example, '*birth_place*', '*birthplace and age*' and '*place birth*' are mapped to '*birthPlace*'. OntoCloud v0.2 includes 1,927 classes, 74 object properties and 101 data properties.

We provided the first implementation of the DBpedia/Wikipedia multi-lingual enrichment research.

4 Related Work

DBpedia[6] focuses on extracting information from Wikipedia and make it usable for the Semantic Web. There are several other projects which have the same goal.

The first project is Yago[7]. Yago extracts information from Wikipedia and WordNet. It concentrates on the category system and the Infoboxes of Wikipedia and combines this information with the taxonomy of WordNet.

Another approach is Semantic MediaWiki [8] [9]. It is an extension for MediaWiki, the system used for Wikipedia. This extension allows you to add structured data into Wikis by using a specific syntax.

The third project is Freebase, an online database of structured data. Users can edit this database in a similar way as Wikipedia can be edited.

In the area of cross-language data fusion, another project has been launched [10]. The goal is to extract Infobox data from multiple Wikipedia editions and fusing the extracted data among editions. To increase the quality of articles, missing data in one edition will be complemented by data from other editions. If a value exists more than once, the property which is most likely correct will be selected.

The DBpedia ontology has been created manually based on the most commonly used Infoboxes within Wikipedia. Kylin Ontology Generator[11] is an autonomous system for refining such an ontology. To achieve this, the system combines Wikipedia Infoboxes with WordNet using statistical-relational learning.

The last project is CoreOnto. It is the research project about IT ontology infrastructure and service technology development [9]. There are several components and solutions for semi-automated ontology construction. One of them is the CAT2ISA[12] which is a toolkit to extract isa/instanceOf relation from category structure. It supports not only lexical patterns, but it also analyze other category links related to the given category link to determine whether the given category link is isa/instanceOf relation or not.

[9] CoreOnto http://ontocore.org

5 Future Work and Conclusion

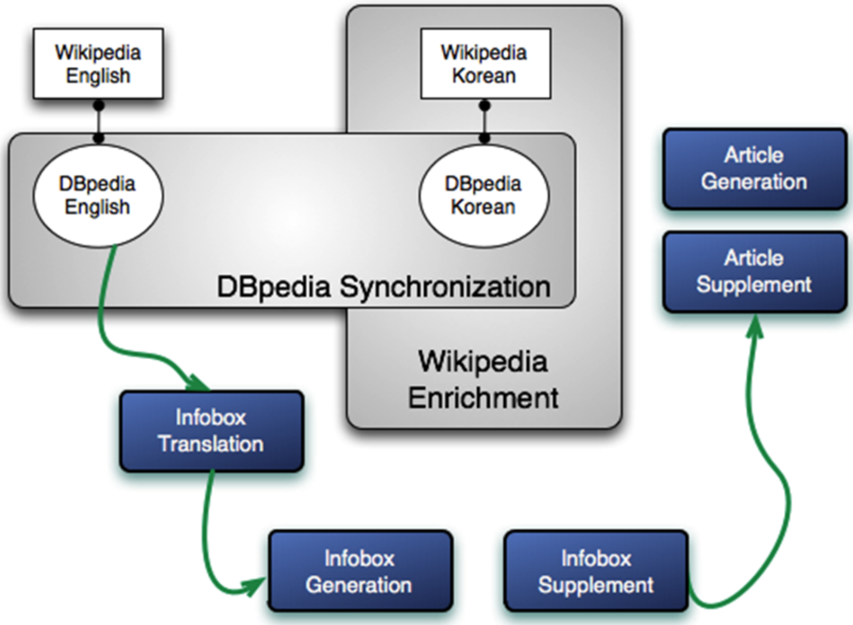


Fig. 4. English-Korean Enrichment in DBpedia/Wikipedia

As future work, it is planned to support more extractors for the Korean language and improve the quality of the extracted datasets. The support for Yago and WordNet could be covered by DBpedia-OntoCloud. The OntoCloud has been linked to WordNet where OntoCloud is an ontology transformed from the templates of infobox (English). To make the dataset accessible for everybody a server will be set up at the SWRC[10] at KAIST[11].

Because encoding for Korean results in unreadable strings for human beings the idea has been raised to use IRI's instead of URI's. It is still uncertain if all tools of the tool chain can handle IRI's. Nevertheless it is already possible to extract the data with IRI's if desired. These triples contain characters that are not valid in XML. The number of triples with such characters is only 48 and can be ignored for now. We also plan to set up a Virtuoso Server to query the Korean DBpedia over SPARQL.

The results from the translated Infoboxes should be evaluated precisely and improved afterwards. After the verification of the triples, the Korean DBpedia

[10] http://swrc.kaist.ac.kr
[11] Korea Advanced Institute of Science and Technology: http://www.kaist.ac.kr

can be updated. This will be helpful to guarantee that the same information can be recognized in different languages.

The concept shown in Fig. 4 describes the information enrichment process within the Wikipedia and DBpedia. As described earlier, we first synchronized two different language versions of DBpedia using translation. After that we can create infoboxes using translated values. In the final study of this project, we will generate new sentences for Wikipedia using newly added data from DBpedia. These sentences will be published in the appropriate Wikipedia articles. It can help authors to edit articles and to create infoboxes when a new article is created. The system can support the authors by suggesting the right template.

References

1. Giles, G. Internet encyclopedias go head to head. Nature, 438 (2005), 900-901, 2005
2. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann, DBpedia - A crystallization point for the Web of Data. Journal of Web Semantics, 7 (3), pp. 154-165, 2009
3. Orri Erling and Ivan Mikhailov. RDF support in the Virtuoso DBMS. Volume P-113 of GI-Edition - Lecture Notes in Informatics (LNI), ISSN 1617-5468, Bonner Köllen Verlag, 2007.
4. Sören Auer and Jens Lehmann. W hat have innsbruck and leipzig in common? Extracting semantics from wiki content. In Enricho Fanconi, Michael Kifer, and Wolfgang May, editors, ESWC, volume 4519 of LNCS, pages 503-517. 2007.
5. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. LNCS, ISSN 1611-3349, Springer Berlin/Heidelberg, 2007.
6. Sebastian Hellmann, Claus Stadler, Jens Lehmann, Sören Auer. DBpedia Live Extraction. Universitat Leipzig (2008), http://www.informatik.uni-leipzig.de/ auer/publication/dbpedia-live-extraction.pdf
7. Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum. YAGO: A Large Ontology from Wikipedia and WordNet. Web Semantics: Science, Services and Agents on the World Wide Web, Volume 6, Issue 3, Pages 203-217, 2008
8. Markus Krötzsch, Denny Vrandecic, and Max Völkel. Wikipedia and the Semantic Web – The Missing Links. In Jakob Voss and Andrew Lih, editors, Proceedings of Wikimania, 2005
9. Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic Wikipedia. WWW 2006, pages 585-594, ACM, 2006.
10. Eugenio Tacchini, Andreas Schultz, Christian Bizer: Experiments with Wikipedia Cross-Language Data Fusion. 5th Workshop on Scripting and Development for the Semantic Web (SFSW2009), Crete, 2009.
11. Fei Wu, Daniel S. Weld: Automatically Refining the Wikipedia Inforbox Ontology. International World Wide Web Conference, Beijing, 2008
12. DongHyun Choi and Key-Sun Choi: Incremental Summarization using Taxonomy, KCAP 2009