

DBpedia - A Linked Data Hub and Data Source for Web and Enterprise Applications

Georgi Kobilarov
Freie Universität Berlin
Garystr. 21
D-14195 Berlin, Germany
georgi.kobilarov@fu-berlin.de

Sören Auer
Universität Leipzig
Johannisgasse 26
D-04103 Leipzig, Germany
auer@uni-leipzig.de

Christian Bizer
Freie Universität Berlin
Garystr. 21
D-14195 Berlin, Germany
chris@bizer.de

Jens Lehmann
Universität Leipzig
Johannisgasse 26
D-04103 Leipzig, Germany
lehmann@informatik.uni-leipzig.de

ABSTRACT

The DBpedia project has extracted a rich knowledge base from Wikipedia and serves this knowledge base as Linked Data on the Web. DBpedia's knowledge base currently provides 274 million pieces of information about 2.6 million concepts. As DBpedia covers a wide range of domains and has a high degree of conceptual overlap with various open-license datasets that are already available on the Web, an increasing number of data publishers has started to set data links from their data sources to DBpedia, making DBpedia one of the central interlinking hubs of the emerging Web of Data. This paper gives an overview about the DBpedia project and describes how application developers can make use of DBpedia knowledge within their applications.

Categories and Subject Descriptors

H.4.m [Information Systems]: Miscellaneous

Keywords

Web of Data, Linked Data, Knowledge Extraction, Wikipedia, DBpedia

1. INTRODUCTION

Knowledge bases are playing an increasingly important role in enhancing the intelligence of Web and enterprise search and in supporting information integration. Today, most knowledge bases cover only specific domains, are created by relatively small groups of knowledge engineers, and are very cost intensive to keep up-to-date as domains change. At the same time, Wikipedia has grown into one of the central knowledge sources of mankind, maintained by thousands of contributors. The DBpedia project [1] leverages this gigantic source of knowledge by extracting structured information from Wikipedia and by making this information accessible on the Web.

Copyright is held by the author/owner(s).
WWW2009, April 20-24, 2009, Madrid, Spain.

The DBpedia knowledge base currently describes more than 2.6 million things, including at least 213,000 persons, 328,000 places, 57,000 music albums, 36,000 films, 20,000 companies. The knowledge base consists of 274 million pieces of information (RDF triples). It features labels and short abstracts for these things in 15 different languages; 609,000 links to images and 3,150,000 links to external web pages; 4,878,100 external links into other RDF datasets. Entities are classified in 4 concept hierarchies: The manually build DBpedia ontology, the YAGO [6] ontology, the UMBEL¹ ontology and a SKOS representation of the Wikipedia category system. The DBpedia knowledge base has several advantages over existing knowledge bases: It covers many domains, it represents real community agreement, it automatically evolves as Wikipedia changes, and it is truly multilingual.

This paper is structured as follows: We give an overview of the DBpedia extraction framework and describe how Web applications can access the DBpedia knowledge base. Afterwards, we describe three use cases of the DBpedia knowledge base and its concept identifiers: Knowledge source for web applications; interlinking hub to connect data sources, and vocabulary for annotating web documents.

2. DBPEDIA EXTRACTION FRAMEWORK

While Wikipedia articles consist mostly of free text, they also contain various types of structured information, such as infobox templates, categorisation information, images, geo coordinates, links to external Web pages and other Wikipedia articles, disambiguation information, redirects and cross-language links. The DBpedia extraction framework extracts these different kinds of information and turns them into RDF data.

All entities in DBpedia are assigned a unique URI of the form <http://dbpedia.org/resource/Name>, where *Name* is taken from the URL of the source Wikipedia article, which has the form <http://en.wikipedia.org/wiki/Name>.

¹<http://umbel.org>

The type of wiki contents that are most valuable for the DBpedia extraction are Wikipedia infoboxes. Infoboxes contain attribute value pairs and are used to display an article's most relevant facts as a table at the top right-hand side of the corresponding Wikipedia page. Wikipedia's infobox template system has evolved over time without central coordination. Therefore, there is a lack of uniformity of infoboxes. Different templates use different names for the same attribute (e.g. `birthplace` and `placeofbirth`). While the first version of our infobox extractor used a generic method to turn property value pairs into triples and hence struggled with the different names of attributes, our new mapping-based extractor aims to solve that problem by introducing a central DBpedia ontology and mappings between templates and the ontology.

This ontology was created by manually arranging the 350 most commonly used infobox templates within the English edition of Wikipedia into a subsumption hierarchy consisting of 170 classes and then mapping 2300 attributes from within these templates to 720 ontology properties. The property mappings define fine-grained rules on how to parse infobox values and define target datatypes, which help the parsers to process values.

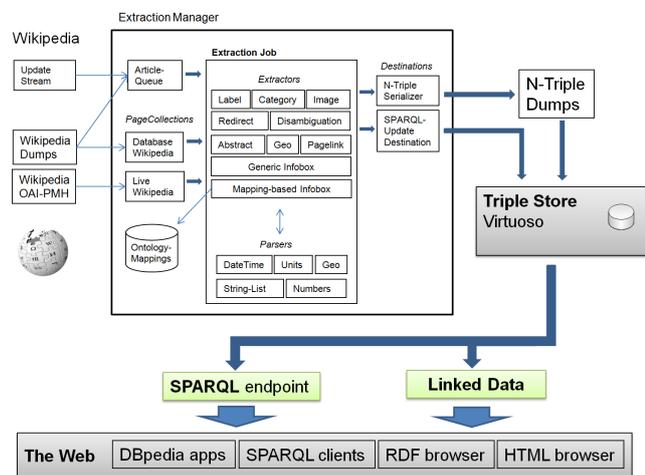


Figure 1: Overview of DBpedia components.

Figure 1 gives an overview of the open-source DBpedia extraction framework. The main components of the framework are: *PageCollections* which are an abstraction of local or remote sources of Wikipedia articles, *Destinations* that store or serialize extracted RDF triples, *Extractors* which turn a specific type of wiki markup into triples, *Parsers* which support the extractors by determining datatypes, conversion values between different units and splitting markup into lists. *ExtractionJobs* group a page collection, extractions and a destination into a workflow. The core of the framework is the *Extraction Manager* which manages the process of passing Wikipedia articles to the extractors and delivers their output to the destination.

3. ACCESSING DBPEDIA OVER THE WEB

In order to fulfill the requirements of different client applications, we serve the DBpedia knowledge through four access mechanisms:

Linked Data. DBpedia URIs be dereferenced over the Web according to the Linked Data principles [2, 3]. DBpedia resource identifiers (such as <http://dbpedia.org/resource/Berlin>) are set up to return (a) RDF descriptions when accessed by Semantic Web agents (such as data browsers or crawlers of Semantic Web search engines), and (b) a simple HTML view of the same information to traditional Web browsers. HTTP content negotiation is used to deliver the appropriate format.

SPARQL Endpoint. We provide a SPARQL endpoint for querying the DBpedia knowledge base. Client applications can send queries over the SPARQL protocol to this endpoint at <http://dbpedia.org/sparql>.

RDF Dumps. N-Triple serializations of the datasets are available for download at the DBpedia website at <http://wiki.dbpedia.org/Downloads32>.

Lookup Index. In order to make it easy for Linked Data publishers to find DBpedia resource URIs to link to, we provide a lookup service that proposes DBpedia URIs for a given label. The Web service is available at <http://lookup.dbpedia.org/api/search.asmx>.

4. USE CASES

This section describes three use cases of the DBpedia knowledge base and its concept identifiers.

4.1 Data Source

The DBpedia knowledge base is served on the Web under the terms of the GNU Free Documentation License. Application can therefore query the knowledge and use the query results, including labels and abstracts in 15 languages, for their purposes. Did you ever need a list and abstracts about 'Dutch cities over 200 meters altitude', 'Italian musicians from the 18th century', 'episodes of the HBO television show "The Sopranos"' or 'Software developed by an organisation founded in California by a person born in a European country in the 1960s' for your application? The DBpedia knowledge base can provide them for you. More sample SPARQL queries can be found on the DBpedia wiki at <http://wiki.dbpedia.org>.

4.2 Interlinking Hub

Linked Data [2, 3] has become increasingly popular as a lightweight approach to publishing and connecting data on the Web. Over the last year, an increasing number of data publishers have started to set data links to DBpedia concepts, making DBpedia a central interlinking hub for the emerging Web of data. Currently, the Web of interlinked data sources around DBpedia provides around 4.5 billion pieces of information and covers domains such as geographic information, people, companies, films, music, genes, drugs, books, and scientific publications².

A major advantage of using DBpedia as linking hub is that it contains semantic relations bridging different domains. This way specialized domain-specific datasets linked to DBpedia can be leveraged in cross domain (and cross dataset) queries. Figure 2 shows the cloud of interlinked data sources

²<http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

