Web Semantics: Science, Services and Agents on the World Wide Web xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

DBpedia - A crystallization point for the Web of Data

Christian Bizer[a,*], Jens Lehmann[b,*], Georgi Kobilarov[a], Sören Auer[b], Christian Becker[a], Richard Cyganiak[c], Sebastian Hellmann[b]

[a] Freie Universität Berlin, Web-based Systems Group, Garystr. 21, D-14195 Berlin, Germany
[b] Universitdt Leipzig, Department of Computer Science, Johannisgasse 26, D-04103 Leipzig, Germany
[c] Digital Enterprise Research Institute, National University of Ireland, Lower Dangan, Galway, Ireland

ARTICLE INFO

Article history:
Received 28 January 2009
Received in revised form 25 May 2009
Accepted 1 July 2009
Available online xxx

Keywords:
Web of Data
Linked Data
Knowledge extraction
Wikipedia
RDF

ABSTRACT

The DBpedia project is a community effort to extract structured information from Wikipedia and to make this information accessible on the Web. The resulting DBpedia knowledge base currently describes over 2.6 million entities. For each of these entities, DBpedia defines a globally unique identifier that can be dereferenced over the Web into a rich RDF description of the entity, including human-readable definitions in 30 languages, relationships to other resources, classifications in four concept hierarchies, various facts as well as data-level links to other Web data sources describing the entity. Over the last year, an increasing number of data publishers have begun to set data-level links to DBpedia resources, making DBpedia a central interlinking hub for the emerging Web of Data. Currently, the Web of interlinked data sources around DBpedia provides approximately 4.7 billion pieces of information and covers domains such as geographic information, people, companies, films, music, genes, drugs, books, and scientific publications. This article describes the extraction of the DBpedia knowledge base, the current status of interlinking DBpedia with other data sources on the Web, and gives an overview of applications that facilitate the Web of Data around DBpedia.

1. Introduction

Knowledge bases play an increasingly important role in enhancing the intelligence of Web and enterprise search, as well as 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 leverages this gigantic source of knowledge by extracting structured information from Wikipedia and making this information accessible on the Web. The resulting DBpedia knowledge base currently describes more than 2.6 million entities, including 198,000 persons, 328,000 places, 101,000 musical works, 34,000 films, and 20,000 companies. The knowledge base contains 3.1 million links to external web pages; and 4.9 million RDF links into other Web data sources. 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, it is truly multilingual, and it is accessible on the Web.

For each entity, DBpedia defines a globally unique identifier that can be dereferenced according to the Linked Data principles [4,5]. 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 have started to set RDF links from their data sources to DBpedia, making DBpedia one of the central interlinking hubs of the emerging Web of Data. The resulting Web of interlinked data sources around DBpedia contains approximately 4.7 billion RDF triples[1] and covers domains such as geographic information, people, companies, films, music, genes, drugs, books, and scientific publications.

The DBpedia project makes the following contributions to the development of the Web of Data:

- We develop an information extraction framework that converts Wikipedia content into a rich multi-domain knowledge base. By accessing the Wikipedia live article update feed, the DBpedia knowledge base timely reflects the actual state of Wikipedia.

* Corresponding authors.
E-mail addresses: chris@bizer.de (C. Bizer), lehmann@informatik.uni-leipzig.de (J. Lehmann).

[1] http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics

1570-8268/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2009.07.002

A mapping from Wikipedia infobox templates to an ontology increases the data quality.
- We define a Web-dereferenceable identifier for each DBpedia entity. This helps to overcome the problem of missing entity identifiers that has hindered the development of the Web of Data so far and lays the foundation for interlinking data sources on the Web.
- We publish RDF links pointing from DBpedia into other Web data sources and support data publishers in setting links from their data sources to DBpedia. This has resulted in the emergence of a Web of Data around DBpedia.

This article documents the recent progress of the DBpedia effort. It builds upon two earlier publications about the project [1,2]. The article is structured as follows: we give an overview of the DBpedia architecture and knowledge extraction techniques in Section 2. The resulting DBpedia knowledge base is described in Section 3. Section 4 discusses the different access mechanisms that are used to serve DBpedia on the Web. In Section 5, we give an overview of the Web of Data that has developed around DBpedia. We showcase applications that facilitate DBpedia in Section 6 and review related work in Section 7. Section 8 concludes and outlines future work.

2. The DBpedia knowledge extraction framework

Wikipedia articles consist mostly of free text, but also contain various types of structured information in the form of wiki markup. Such information includes infobox templates, categorisation information, images, geo-coordinates, links to external Web pages, disambiguation pages, redirects between pages, and links across different language editions of Wikipedia. The DBpedia project extracts this structured information from Wikipedia and turns it into a rich knowledge base. In this chapter, we give an overview of the DBpedia knowledge extraction framework, and discuss DBpedia's infobox extraction approach in more detail.

2.1. Architecture of the extraction framework

Fig. 1 gives an overview of the DBpedia knowledge 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, converting values between different units and splitting markup into lists. *Extraction Jobs* group a page collection, extractors 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. The Extraction Manager also handles URI management and resolves redirects between articles.

The framework currently consists of 11 extractors which process the following types of Wikipedia content:

- *Labels*. All Wikipedia articles have a title, which is used as an `rdfs:label` for the corresponding DBpedia resource.
- *Abstracts*. We extract a short abstract (first paragraph, represented using `rdfs:comment`) and a long abstract (text before a table of contents, at most 500 words, using the property `dbpedia:abstract`) from each article.
- *Interlanguage links*. We extract links that connect articles about the same topic in different language editions of Wikipedia and use them for assigning labels and abstracts in different languages to DBpedia resources.
- *Images*. Links pointing at Wikimedia Commons images depicting a resource are extracted and represented using the `foaf:depiction` property.
- *Redirects*. In order to identify synonymous terms, Wikipedia articles can redirect to other articles. We extract these redirects and use them to resolve references between DBpedia resources.
- *Disambiguation*. Wikipedia disambiguation pages explain the different meanings of homonyms. We extract and represent disambiguation links using the predicate `dbpedia:disambiguates`.
- *External links*. Articles contain references to external Web resources which we represent using the DBpedia property `dbpedia:reference`.
- *Pagelinks*. We extract all links between Wikipedia articles and represent them using the `dbpedia:wikilink` property.
- *Homepages*. This extractor obtains links to the homepages of entities such as companies and organisations by looking for the terms *homepage* or *website* within article links (represented using `foaf:homepage`).
- *Categories*. Wikipedia articles are arranged in categories, which we represent using the SKOS vocabulary.[2] Categories become `skos:concepts`; category relations are represented using `skos:broader`.
- *Geo-coordinates*. The geo-extractor expresses coordinates using the Basic Geo (WGS84 lat/long) Vocabulary[3] and the GeoRSS Simple encoding of the W3C Geospatial Vocabulary.[4] The former expresses latitude and longitude components as separate facts, which allows for simple areal filtering in SPARQL queries.

The DBpedia extraction framework is currently set up to realize two workflows: a regular, dump-based extraction and the live extraction.

2.1.1. Dump-based extraction

The Wikimedia Foundation publishes SQL dumps of all Wikipedia editions on a monthly basis. We regularly update the DBpedia knowledge base with the dumps of 30 Wikipedia editions. The dump-based workflow uses the *DatabaseWikipedia page collection* as the source of article texts and the N-Triples serializer as the output destination. The resulting knowledge base is made available as Linked Data, for download, and via DBpedia's main SPARQL endpoint (cf. Section 4).

2.1.2. Live extraction

The Wikimedia Foundation has given the DBpedia project access to the *Wikipedia OAI-PMH live feed* that instantly reports all Wikipedia changes. The live extraction workflow uses this update stream to extract new RDF whenever a Wikipedia article is changed. The text of these articles is accessed via the *LiveWikipedia page collection*, which obtains the current article version encoded according to the OAI-PMH protocol. The *SPARQL-Update Destination* deletes existing and inserts new triples into a separate triple store. According to our measurements, about 1.4 article pages are updated each second on Wikipedia. The framework can handle up to 8.8 pages per second on a 2.4 GHz dual-core machine (this includes consumption from the stream, extraction, diffing and loading the triples into a Virtuoso triple store). The time lag for DBpedia to reflect Wikipedia changes lies between 1 or 2 min. The bottleneck here is the update stream, since changes normally need more than 1 min to arrive from Wikipedia. More information about the live extraction is found at http://en.wikipedia.org/wiki/User:DBpedia.

[2] http://www.w3.org/2004/02/skos/
[3] http://www.w3.org/2003/01/geo/
[4] http://www.w3.org/2005/Incubator/geo/XGR-geo/

Please cite this article in press as: C. Bizer, et al., DBpedia - A crystallization point for the Web of Data, Web Semantics: Sci. Serv. Agents World Wide Web (2009), doi:10.1016/j.websem.2009.07.002

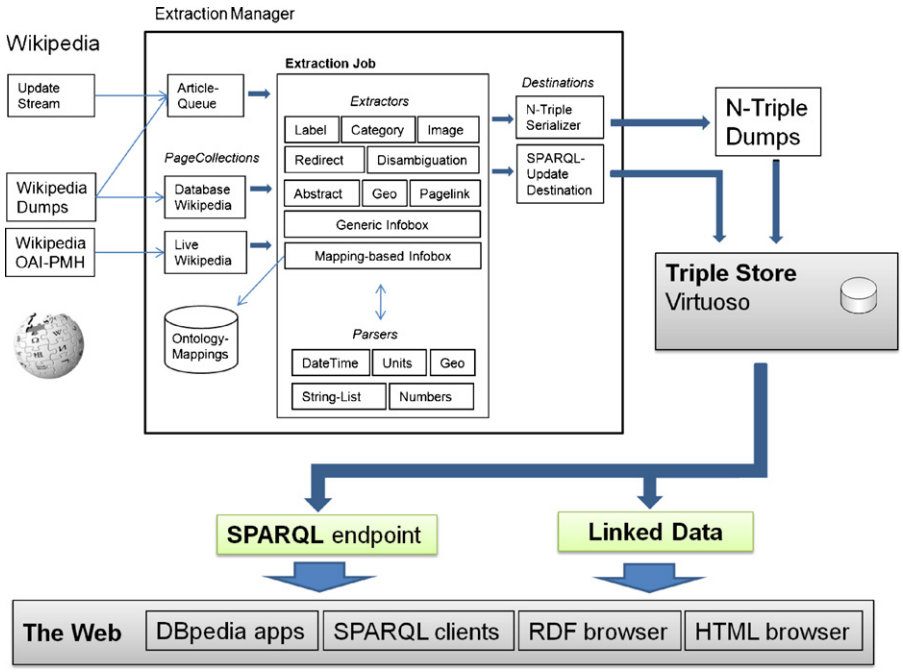

Fig. 1. Overview of DBpedia components.

{{ Infobox Actor
| birthname = Thomas Jeffrey Hanks
| birthdate = {{birth date and age|1956|7|9}}
| birthplace = [[Concord, California|Concord]],
[[California]]
| yearsactive = 1979 - present
| occupation = Actor, producer, director,
[[voice over artist]],
writer, speaker

Fig. 2. Infobox Tom Hanks.

2.2. Generic versus mapping-based infobox extraction

The type of wiki contents that is most valuable for the DBpedia extraction are Wikipedia infoboxes. Infoboxes display an article's most relevant facts as a table of attribute-value pairs on the top right-hand side of the Wikipedia page. Figs. 2 and 3 show excerpts of the wiki markup behind the infoboxes describing Tom Hanks and Andre Agassi. Wikipedia's infobox template system has evolved over time without central coordination. Different communities use different templates to describe the same type of things (e.g. `inf obox_city_japan`, `inf obox_swiss_town` and `inf obox_town_de`). Different templates use different names for the same attribute (e.g. `birthplace` and `placeofbirth`). As many Wikipedia editors do not strictly follow the recommendations given on the page that describes a template, attribute values are expressed using a wide range of different formats and units of measurement. The DBpedia project has decided to deal with this situation by using two different extraction approaches in parallel: a generic approach which aims at wide coverage and a mapping-based approach which aims at high data quality.

2.2.1. Generic infobox extraction

The generic infobox extraction algorithm, which is described in detail in [2], processes all infoboxes within a Wikipedia article. It creates triples from the infobox data in the following manner: the corresponding DBpedia URI of the Wikipedia article is used as subject. The predicate URI is created by concatenating the namespace fragment http://dbpedia.org/property/ and the name of the infobox attribute. Objects are created from the attribute value. Property values are post-processed in order to generate suitable URI references or literal values. This includes recognizing MediaWiki links, detecting lists, and using units as datatypes. MediaWiki templates may be nested, which we handle through a blanknode creation algorithm. The advantage of the generic extraction is its complete coverage of all infoboxes and infobox attributes. The main disadvantage is that synonymous attribute names are not resolved, which makes writing queries against generic infobox data rather cumbersome. As Wikipedia attributes do not have explicitly defined datatypes, a further problem is the relatively high error rate of the heuristics that are used to determine the datatypes of attribute values.

{{ Infobox Tennis Player
| country = United States
| playername = Andre Agassi
| residence = [[Las Vegas metropolitan area|LasVegas]],
[[Nevada]], United States
| datebirth = {{birth date and age| mf=yes|1970|5|29}}
| placebirth = [[Las Vegas, Nevada]], United States
| height = {{convert|1.80|m|ftin|abbr=on}}
| weight = {{convert|177|lb|kg|abbr=on}}

Fig. 3. Infobox Andre Agassi.

Please cite this article in press as: C. Bizer, et al., DBpedia - A crystallization point for the Web of Data, Web Semantics: Sci. Serv. Agents World Wide Web (2009), doi:10.1016/j.websem.2009.07.002

2.2.2. Mapping-based infobox extraction

In order to overcome the problems of synonymous attribute names and multiple templates being used for the same type of things, we mapped Wikipedia templates to an 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 2350 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 attribute values. For instance, if a mapping defines the target datatype to be a list of links, the parser will ignore additional text that might be present in the attribute value. The ontology currently uses 55 different datatypes. Deviant units of measurement are normalized to one of these datatypes. Instance data within the infobox ontology is therefore cleaner and better structured than data that is generated using the generic extraction algorithm. The disadvantage of the mapping-based approach is that it currently covers only 350 Wikipedia templates; therefore it only provides data about 843,000 entities compared to 1,462,000 entities that are covered by the generic approach. While the ontology is currently relatively simple, we plan to extend it further, e.g. with class disjointness axioms and inverse properties. The main purpose of such extensions will be to allow consistency checks in DBpedia and use inferences when answering SPARQL queries. The members of the DBpedia team will not be able to extend the ontology, the mappings and the parser rules to cover all Wikipedia infoboxes, due to the size of the task and the knowledge required to map templates from exotic domains. We are therefore working on methods to crowd-source this task. We are currently developing an approach to integrate the DBpedia ontology itself back into Wikipedia. Ontology definitions related to a certain infobox templates will be represented themselves as infoboxes on the corresponding template definition page. Combined with the live extraction, the wider Wikipedia community would thus have a powerful tool for extending and refining of both—the ontology and the infobox mappings.

3. The DBpedia knowledge base

The DBpedia knowledge base currently consists of around 274 million RDF triples, which have been extracted from the English, German, French, Spanish, Italian, Portuguese, Polish, Swedish, Dutch, Japanese, Chinese, Russian, Finnish, Norwegian, Catalan, Ukranian, Turkish, Czech, Hungarian, Romanian, Volapiik, Esperanto, Danish, Slovak, Indonesian, Arabic, Korean, Hebrew, Lithuanian, Vietnamese, Slovenian, Serbian, Bulgarian, Estonian, and Welsh versions of Wikipedia. The knowledge base describes more than 2.6 million entities. It features labels and short abstracts in 30 different languages; 609,000 links to images; 3,150,000 links to external web pages; 415,000 Wikipedia categories, and 286,000 YAGO categories.

Table 1 gives an overview of common DBpedia classes, and shows the number of instances and some example properties for each class. In the following, we describe the structure of the DBpedia knowledge base, explain how identifiers are built and compare the four classification schemata that are offered by DBpedia.

3.1. Identifying entities

DBpedia uses English article names for creating identifiers. Information from other language versions of Wikipedia is mapped to these identifiers by bi-directionally evaluating the interlanguage links between Wikipedia articles. Resources are assigned a URI according to the pattern 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. This yields certain beneficial properties:

- DBpedia URIs cover a wide range of encyclopedic topics.
- They are defined by community consensus.
- There are clear policies in place for their management.
- A extensive textual definition of the entity is available at a well-known Web location (the Wikipedia page).

3.2. Classifying entities

DBpedia entities are classified within four classification schemata in order to fulfill different application requirements. We compare these schemata below:

Wikipedia Categories. DBpedia contains a SKOS representation of the Wikipedia category system. The category system consists of 415,000 categories. The main advantage of the category system is that it is collaboratively extended and kept up-to-date by thousands of Wikipedia editors. A disadvantage is that categories do not form a proper topical hierarchy, as there are cycles in the category system and as categories often only represent a rather loose relatedness between articles.

YAGO. The YAGO classification schema consists of 286,000 classes which form a deep subsumption hierarchy. The schema was created by mapping Wikipedia leaf categories, i.e. those not having sub categories, to WordNet synsets. Details of the mapping algorithm are described in [19]. Characteristics of the YAGO hierarchy are its deepness and the encoding of much information in one class (e.g. the class "MultinationalCompaniesHeadquarteredInTheNetherlands"). While YAGO achieves a high accuracy in general, there are a few errors and omissions (e.g. the mentioned class is not a subclass of "MultinationalCompanies") due to its automatic generation. We jointly developed a script that

Table 1
Common DBpedia classes with the number of their instances and example properties.

Ontology class	Instances	Example properties
Person	198,056	name, birthdate, birthplace, employer, spouse
Artist	54,262	activeyears, awards, occupation, genre
Actor	26,009	academyaward, goldenglobeaward, activeyears
MusicalArtist	19,535	genre, instrument, label, voiceType
Athlete	74,832	current Team, currentPosition, currentNumber
Politician	12,874	predecessor, successor, party
Place	247,507	lat, long
Building	23,304	architect, location, openingdate, style
Airport	7,971	location, owner, IATA, lat, long
Bridge	1,420	crosses, mainspan, openingdate, length
Skyscraper	2,028	developer, engineer, height, architect, cost
PopulatedPlace	181,847	foundingdate, language, area, population
River	10,797	sourceMountain, length, mouth, maxDepth
Organisation	91,275	location, foundationdate, keyperson
Band	14,952	currentMembers, foundation, homeTown, label
Company	20,173	industry, products, netincome, revenue
Educ.Institution	21,052	dean, director, graduates, staff, students
Work	189,620	author, genre, language
Book	15,677	isbn, publisher, pages, author, mediatype
Film	34,680	director, producer, starring, budget, released
MusicalWork	101,985	runtime, artist, label, producer
Album	74,055	artist, label, genre, runtime, producer, cover
Single	24,597	album, format, releaseDate, band, runtime
Software	5,652	developer, language, platform, license
TelevisionShow	10,169	network, producer, episodenumber, theme

Please cite this article in press as: C. Bizer, et al., DBpedia - A crystallization point for the Web of Data, Web Semantics: Sci. Serv. Agents World Wide Web (2009), doi:10.1016/j.websem.2009.07.002

Table 2
Comparison of the generic infobox, mapping-based infobox and pagelinks datasets.

	Described entities	Mio. triples	Unique properties	Triples/property	Triples/entity
Generic extraction	1,462,108	26.0	38,659	673.7	17.81
Mapping-based extraction	843,169	7.0	720	9722.2	8.34
Pagelinks	2,853,315	70.2	1	70.2mio	24.61

Table 3
Comparison of the graph structure of the generic infobox, mapping-based infobox and pagelinks datasets.

	Connected entities	Mio. triples	Unique properties	Indegree		Cluster coefficient
				Max	Avg	
Generic extraction	1,029,712	5.6	9911	105,840	8.76	0.1336
Mapping-based extraction	627,941	2.9	340	65,387	11.03	0.1037
Pagelinks	2,796,401	46.2	1	190,995	19.15	0.1696

assigns YAGO classes to DBpedia entities. The script is available at the YAGO download page.[5]

UMBEL. The Upper Mapping and Binding Exchange Layer (UMBEL) is a lightweight ontology that has been created for interlinking Web content and data. UMBEL was derived from *OpenCyc* and consists of 20,000 classes. OpenCyc classes in turn are partially derived from Cyc collections, which are based on WordNet synsets. Since YAGO also uses WordNet synsets and is based on Wikipedia, a mapping from OpenCyc classes to DBpedia can be derived via UMBEL.[6] The classification is maintained by the UMBEL project itself and details about its generation process can be found at the UMBEL website.[7]

DBpedia ontology. The DBpedia ontology consists of 170 classes that form a shallow subsumption hierarchy. It includes 720 properties with domain and range definitions. The ontology was manually created from the most commonly used infobox templates within the English edition of Wikipedia. The ontology is used as target schema by the mapping-based infobox extraction described in Section 2.2. The left column in Table 1 displays a part of the class hierarchy of the DBpedia ontology.

3.3. Describing entities

Every DBpedia entity is described by a set of general properties and a set of infobox-specific properties, if the corresponding English Wikipedia article contains an infobox. The general properties include a label, a short and a long English abstract, a link to the corresponding Wikipedia article (if available) geo-coordinates, a link to an image depicting the entity, links to external Web pages, and links to related DBpedia entities. If an entity exists in multiple language versions of Wikipedia, then short and long abstracts within these languages and links to the different language Wikipedia articles are added to the description.

Infobox-specific properties which result from the generic extraction are defined in the http://dbpedia.org/property/ namespace. Properties resulting from the mapping-based infobox extraction are defined in the namespace http://dbpedia.org/ontology/.

Table 2 compares the datasets that result from generic infobox extraction, the mapping-based infobox extraction and from the extraction of links between Wikipedia pages (all numbers are for DBpedia release 3.2, English version). DBpedia contains generic infobox data for 1,462,000 resources compared to 843,000 resources that are covered by the mapping-based approach. There are links between 2,853,315 Wikipedia pages. This number of pages is higher than the number of entities in DBpedia (2.6 million) as there are additional pages for lists and templates in Wikipedia. The mapping-based dataset contains 720 different properties compared to 38,659 different properties that are used within the generic dataset (including many synonymous properties). The pagelinks dataset uses a single property to represent the untyped links between Wikipedia pages.

In a next step, we measured characteristics of the RDF graph that connects DBpedia entities. For doing this, we removed all triples from the datasets that did not point at a DBpedia entity, including all literal triples, all external links and all dead links. The size of and number of 'link' properties within these reduced datasets is listed in Table 3. Removing the triples showed, that the percentage of properties pointing to other DBpedia entities is much higher in the mapping-based dataset (53%) compared to generic dataset (25.6%). We calculated the average node indegree as the sum of all inbound edges divided by the number of objects, which had at least one inbound edge from the dataset. This allows to analyse the indegree separately from the coverage or the size of the dataset. The entity with the highest indegree within all three datasets is *United States*. As shown in Fig. 4, the node indegrees follow a power-law distribution in all datasets which is a typical characteristic of small world networks [18]. The clustering coefficient given in the last column of Table 3 was calculated as the number of existing connections between neighbors of a node, divided by possible connections in a directed graph ($k^*(k-1)$, k = number of node neighbors) and averaged over all nodes. The mapping-based approach has a slightly lower clustering coefficient because of its lower coverage.

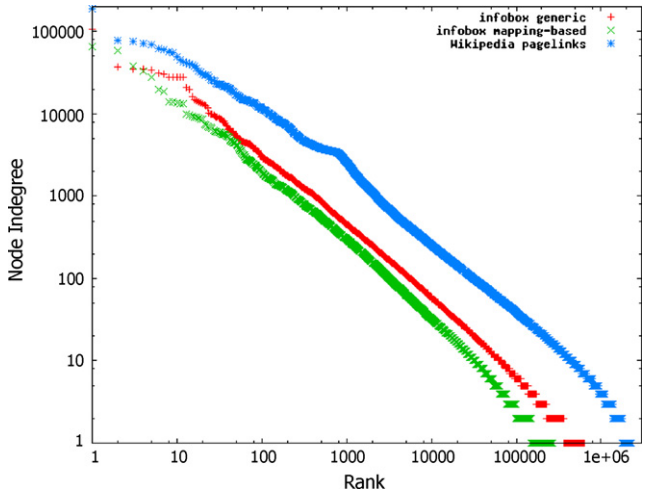

Fig. 4. Comparison of the generic infobox, mapping-based infobox and pagelinks datasets in terms of node indegree versus rank.

[5] http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html
[6] http://fgiasson.com/blog/index.php/2008/09/04/exploding-dbpedias-domain-using-umbel/
[7] http://www.umbel.org/

Please cite this article in press as: C. Bizer, et al., DBpedia - A crystallization point for the Web of Data, Web Semantics: Sci. Serv. Agents World Wide Web (2009), doi:10.1016/j.websem.2009.07.002

<http://dbpedia.org/resource/Spain> owl:sameAs

http://www4.wiwiss.fu-berlin.de/eurostat/resource/countries/Espa%C3%B1a ;

http://rdf.freebase.com/ns/guid.9202a8c04000641f8000000000034e30 ;

http://www4.wiwiss.fu-berlin.de/factbook/resource/Spain ;

http://sw.opencyc.org/2008/06/10/concept/Mx4rvVjowpwpEbGdrcN5Y29ycA .

<http://data.semanticweb.org/conference/eswc/2008/paper/356>

swc:hasTopic <http://dbpedia.org/resource/Data_integration> .

Fig. 5. Example RDF links connecting the DBpedia entity *Spain* with additional information from other data sources, and showing how the DBpedia identifier *Data Integration* is used to annotate the topic of a conference paper.

4. Accessing the DBpedia knowledge base over the web

DBpedia is served on the Web under the terms of the GNU Free Documentation License. In order to fulfill the requirements of different client applications, we provide the DBpedia knowledge base through four access mechanisms:

Linked Data. Is a method of publishing RDF data on the Web that relies on HTTP URIs as resource identifiers and the HTTP protocol to retrieve resource descriptions [4,5]. 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 the endpoint at http://dbpedia.org/sparql. In addition to standard SPARQL, the endpoint supports several extensions of the query language that have proved useful for developing client applications, such as full text search over selected RDF predicates, and aggregate functions, notably COUNT ( ). To protect the service from overload, limits on query complexity and result size are in place. The endpoint is hosted using Virtuoso Universal Server.[8]

RDF dumps. We have sliced the DBpedia knowledge base by triple predicate into several parts and offer N-Triple serialisations of these parts for download on the DBpedia website.[9] In addition to the knowledge base that is served as Linked Data and via the SPARQL endpoint, the download page also offers infobox datasets that have been extracted from Wikipedia editions in 29 languages other than English. These datasets can be used as foundation for fusing knowledge between Wikipedia editions or to build applications that rely on localized Wikipedia knowledge.

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 based on a Lucene index providing a weighted label lookup, which combines string similarity with a relevance ranking (similar to PageRank) in order to find the most likely matches for a given term. DBpedia lookup is available as a Web service at http://lookup.dbpedia.org/api/search.asmx.

Table 4
Distribution of outgoing RDF links pointing from DBpedia to other datasets.

Data source	No. of links
Freebase	2,400,000
flickr wrappr	1,950,000
WordNet	330,000
GeoNames	85,000
OpenCyc	60,000
UMBEL	20,000
Bio2RDF	25,000
WikiCompany	25,000
MusicBrainz	23,000
Book Mashup	7,000
Project Gutenberg	2,500
DBLP Bibliography	200
CIA World Factbook	200
EuroStat	200

The DBpedia Web interfaces are described using the Semantic Web Crawling Sitemap Extension format.[10] Client applications can use this description to choose the most efficient access mechanism for the task they perform.

5. Interlinked web content

In order to enable DBpedia users to discover further information, the DBpedia knowledge base is interlinked with various other data sources on the Web according to the Linked Data principles [4,5]. The knowledge base currently contains 4.9 million outgoing RDF links [5] that point at complementary information about DBpedia entities, as well as meta-information about media items depicting an entity. Over the last year, an increasing number of data publishers have started to set RDF links to DBpedia entities. These incoming links, together with the outgoing links published by the DBpedia project, make DBpedia one of the central interlinking hubs of the emerging Web of Data. These RDF links lay the foundation for:

Web of Data browsing and crawling. RDF links enable information consumers to navigate from data within one data source to related data within other sources using a Linked Data browser [20,3]. RDF links are also followed by the crawlers of Semantic Web search engines, which provide search and query capabilities over crawled data [7,9,21].

Web Data Fusion and Mashups. As RDF links connect data about an entity within different data sources, they can be used as a basis

[8] http://virtuoso.openlinksw.com
[9] http://wiki.dbpedia.org/Downloads32
[10] http://sw.deri.org/2007/07/sitemapextension/

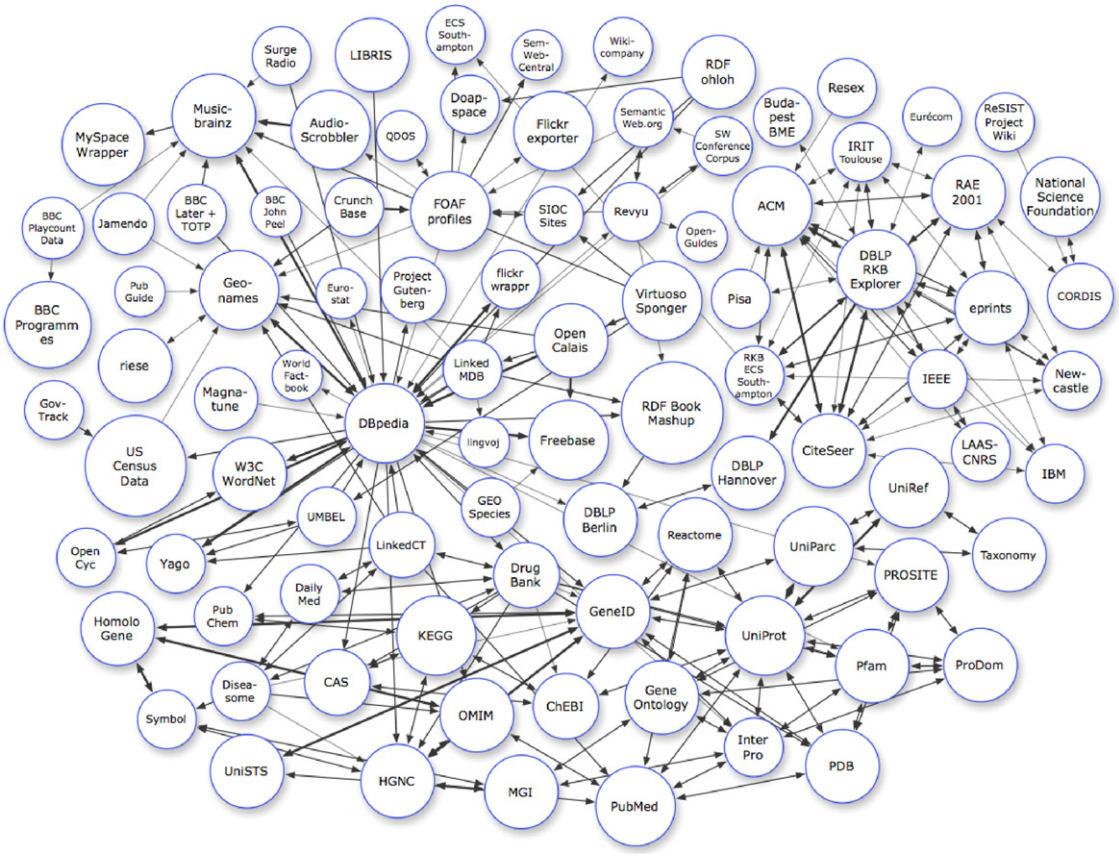

Fig. 6. Data sources that are interlinked with DBpedia.

for fusing data from these sources in order to generate integrated views over multiple sources [16].

Web Content Annotation. DBpedia entity URIs are also used to annotate classic Web content like blog posts or news with topical subjects as well as references to places, companies and people. As the number of sites that use DBpedia URIs for annotation increases, the DBpedia knowledge base could develop into a valuable resource for discovering classic Web content that is related to an entity.

Fig. 5 shows RDF data links that illustrate these use cases. The first four links connect the DBpedia entity *Spain* with complementary data about the country from EuroStat, the CIA World Factbook, Freebase and OpenCyc. Agents can follow these links to retrieve additional information about Spain, which again might contain further deeper links into the data sources. The fifth link illustrates how the DBpedia identifier *Data Integration* is used to annotate the topical subject of a research paper from the European Semantic Web Conference. After this and similar annotations from other sites have been crawled by a search engine, such links enable the discovery of Web content that is related to a topic.

Fig. 6 gives an overview of the data sources that are currently interlinked with DBpedia. Altogether this Web of Data amounts to approximately 4.7 billion RDF triples. Two billion of these triples are served by data sources participating in the *W3C Linking Open Data community project*,[11] an effort to make open-license datasets interoperable on the Web of Data by converting them into RDF and by interlinking them. DBpedia, with its broad topic coverage, intersects with practically all of these datasets and therefore is a useful interlinking hub for such efforts. A second massive source of Linked Data is the *Bio2RDF* project[12] which publishes bio-informatics datasets

Table 5
Data sources publishing RDF links pointing at DBpedia entities.

Data source	Classes
BBC Music	musicians, bands
Bio2RDF	genes, proteins, molecules
CrunchBase	companies
Diseasome	diseases
Faviki	various classes
flickr wrappr	various classes
FOAF	various classes
GeoNames	places
GeoSpecies	species
John Peel	musicians, works
LIBRIS	authors
LinkedCT	intervention, conditions
Linked DrugBank	drugs, diseases
LinkedMDB	films
Lingvoj	languages
OpenCyc	various classes
OpenCalais	locations, people
Surge Radio	musicians, bands
UMBEL	various classes
RDFohloh	programming languages
Revyu	various classes
LODD SIDER	drug side effects
Semantic Web Corpus	various classes

[11] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[12] http://bio2rdf.wiki.sourceforge.net/

Please cite this article in press as: C. Bizer, et al., DBpedia - A crystallization point for the Web of Data, Web Semantics: Sci. Serv. Agents World Wide Web (2009), doi:10.1016/j.websem.2009.07.002

Fig. 7. DBpedia Mobile running on an iPhone 3G and showing a map view of resources in the user's proximity.

as Linked Data on the Web in order to simplify data integration in the genomic field. Altogether, the datasets published by *Bio2RDF* sum up to approximately 2.5 billion triples. A related effort is the *Linking Open Drug Data project*[13] within the *W3C Health Care and Life Sciences interest group* which publishes data about drugs and clinical trials and interlinks published data with the Bio2RDF data cloud as well as with DBpedia.

Table 4 lists the data sources that are reachable from DBpedia by outgoing RDF links.[14] The second column shows the distribution of the 4.9 million outgoing links over the data sources. Using these links, one can, for instance, 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. Outgoing links to ontologies like OpenCyc or UMBEL allow agents to retrieve additional conceptual knowledge which can then be used for reasoning over DBpedia and interlinked data.

Many of the outgoing links were generated based on common identification schemata that are used within Wikipedia and within the external data sources. For instance, Wikipedia articles about books often contain ISBN numbers.

Wikipedia articles about chemical compounds are likely to contain gene, protein and molecule identifiers which are also used by other bio-informatics data sources. For generating the links to GeoNames and MusicBrainz, rule-based approaches that rely on a combination of different properties to match locations (similar name, geo-coordinates, country, administrative division, population)[15] and bands (similar name, similar albums, similar members) [17] are used. In order to maintain the links to other data sources, we plan to employ the *Silk–Link Discovery Framework* [23].

In order to get an overview of the external data sources that currently publish RDF links pointing at DBpedia entities, we analyzed 8 million RDF documents that have been crawled from the Web by the Sindice Semantic Web Search engine [21]. The analysis revealed that there are currently 23 external data sources setting RDF links to DBpedia. Table 5 lists these data sources together with the classes of DBpedia entities that are the targets of the incoming links.

6. Applications facilitated by DBpedia

The DBpedia knowledge base and the Web of Data around DBpedia lay the foundation for a broad range of applications. Section 6.1 describes applications that rely on DBpedia as an interlinking hub to browse and explore the Web of Data. Section 6.2 focuses on applications that use the DBpedia knowledge base to answer complex queries. Section 6.3 gives an overview of applications that use DBpedia entity identifiers for the annotation of Web content.

6.1. Browsing and exploration

As DBpedia is interlinked with various other data sources, DBpedia URIs make good starting points to explore or crawl the Web of Data. Data browsers that can be used to explore the Web of Data include Tabulator [20], Marbles,[16] Disco,[17] and the OpenLink Data Explorer.[18]

6.1.1. DBpedia Mobile

In the following, we describe *DBpedia Mobile* [3], a location-aware client for the Semantic Web that uses DBpedia locations as navigation starting points. DBpedia Mobile[19] allows users to discover, search and publish Linked Data pertaining to their current physical environment using an iPhone and other mobile devices as well as standard web browsers. Based on the current GPS position of a mobile device, DBpedia Mobile renders an interactive map indicating nearby locations from the DBpedia dataset, as shown in Fig. 7. Locations may be labeled in any the 30 languages supported by DBpedia, and are depicted with adequate icons based on a mapping of selected YAGO categories [19]. Starting from this map, the user can explore background information about his surroundings by navigating along data links into other Web data sources: clicking on a resource retrieves web data about the resource, from where RDF links may be followed into other datasets.

DBpedia Mobile is not limited to a fixed set of data sources but may be used to access all data sources that are or will in the future be interlinked with DBpedia or with other data sources that are reachable from DBpedia. This allows interesting navigation paths: from a location, the user may navigate to a person within the DBpedia

[13] http://esw.w3.org/topic/HCLSIG/LODD
[14] For more information about the datasets please refer to http://wiki.dbpedia.org/Interlinking
[15] http://lists.w3.org/Archives/Public/semantic-web/2006Dec/0027.html
[16] http://beckr.org/marbles
[17] http://sites.wiwiss.fu-berlin.de/suhl/bizer/ng4j/disco/
[18] http://ode.openlinksw.com/example.html
[19] http://beckr.org/DBpediaMobile

Please cite this article in press as: C. Bizer, et al., DBpedia - A crystallization point for the Web of Data, Web Semantics: Sci. Serv. Agents World Wide Web (2009), doi:10.1016/j.websem.2009.07.002

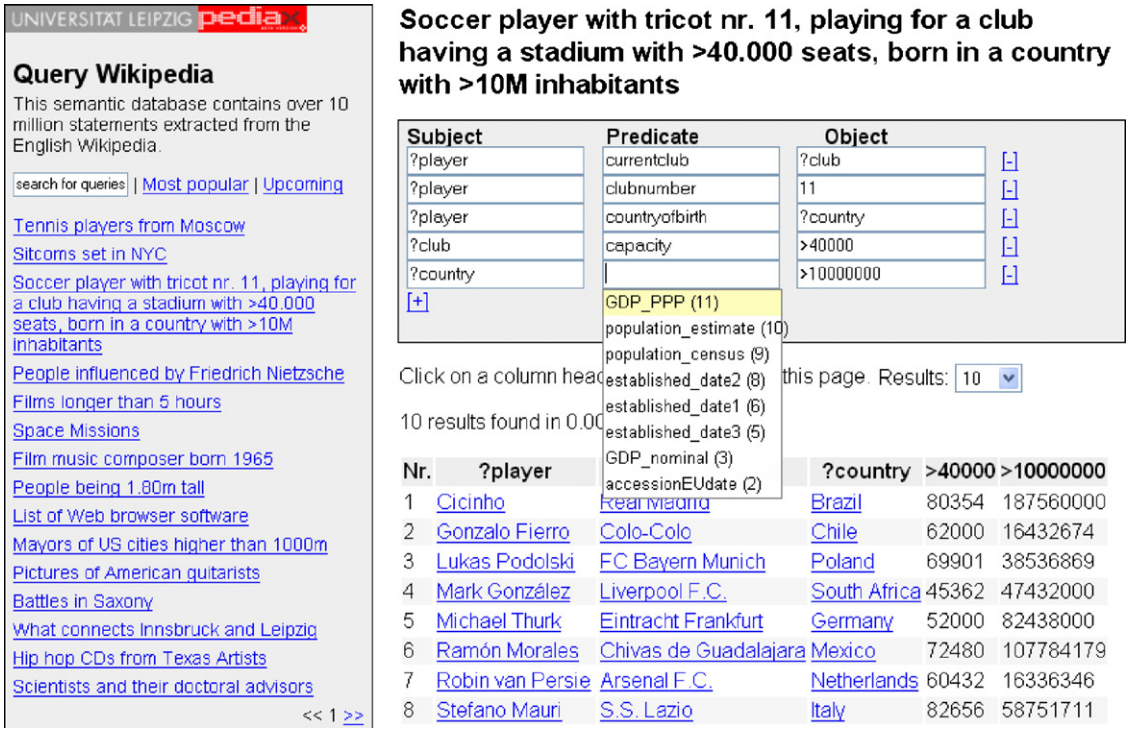

Fig. 8. Form-based DBpedia query builder.

dataset that was born, died or worked at the location. If the person is an author, he may then follow data-level links into the RDF Book Mashup or the Project Gutenberg data sources and explore information about the author's books. If the user is interested in local bands, he may navigate from DBpedia into MusicBrainz and find out more about albums of the bands.

Besides accessing Web data, DBpedia Mobile offers flexible means of filtering using *SPARQL Filters* and enables users to publish their current location, pictures and reviews to the Web of Data so that they can be used by other applications. Instead of simply being tagged with geographical coordinates, published content is interlinked with a nearby DBpedia resource and thus contributes to the overall richness of the geo-spatial Semantic Web.

DBpedia Mobile is based on Marbles, a server-side application that generates entity-centric XHTML views over Web data from several data sources using Fresnel [6] lenses and formats. Prior to rendering a view for a resource, Marbles performs data augmentation, whereby it retrieves interlinked data from the Web and caches retrieved data in an RDF store. This involves dereferencing the resource URI and querying the Sindice [21] and Falcons [7] Semantic Web search engines for related information, as well as Revyu[20] for reviews. Specific predicates found in retrieved data such as `owl:sameAs` and `rdfs:seeAlso` are then followed for up to two levels in order to gain more information about the resource, and to obtain human-friendly resource labels. Marbles employs an `owl:sameAs` inferencer to connect *URI Aliases* [5] between distinct data sources, allowing it to generate unified views of resources.

6.2. Querying and search

The DBpedia knowledge base contains a large amount of general-purpose knowledge and can thus be used to answer quite surprising queries about a wide range of topics.

6.2.1. DBpedia Query Builder

A tool that has been developed to demonstrate these capabilities is the *DBpedia Query Builder*.[21] Fig. 8 shows how the query builder is used to answer a query about soccer players that play for specific clubs and are born in countries with more than 10 million inhabitants. Other example queries are listed in the box on the right-hand side of the screenshot.

Queries are expressed by means of a graph pattern consisting of multiple triple patterns. For each triple pattern three form fields capture variables, identifiers or filters for the subject, predicate and object of a triple. Due to the wide coverage of DBpedia, users can hardly know which properties and identifiers are used in the knowledge base and hence can be used for querying. Consequently, users have to be guided when building queries and reasonable alternatives should be suggested. Therefore, while users type identifier names into one of the form fields, a look-ahead search proposes suitable options. These are obtained not just by looking for matching identifiers but by executing the currently built query using a variable for the currently edited identifier and filtering the results returned for this variable for matches starting with the search string the user supplied. This method ensures that the identifier proposed is really used in conjunction with the graph pattern under construction, and that the query actually returns results.

6.2.2. Relationship Finder

A user interface that can be used to explore the DBpedia knowledge base is the *DBpedia Relationship Finder*.[22] The Relationship Finder allows users to find connections between two different entities in DBpedia. The Relationship Finder user interface initially contains a simple form to enter two entities, as well as a small number of options, and a list of previously saved queries. While typing, the user is offered suggestions for the object he wants to enter. After

[20] http://revyu.com/

[21] http://querybuilder.dbpedia.org/
[22] http://relfinder.dbpedia.org/

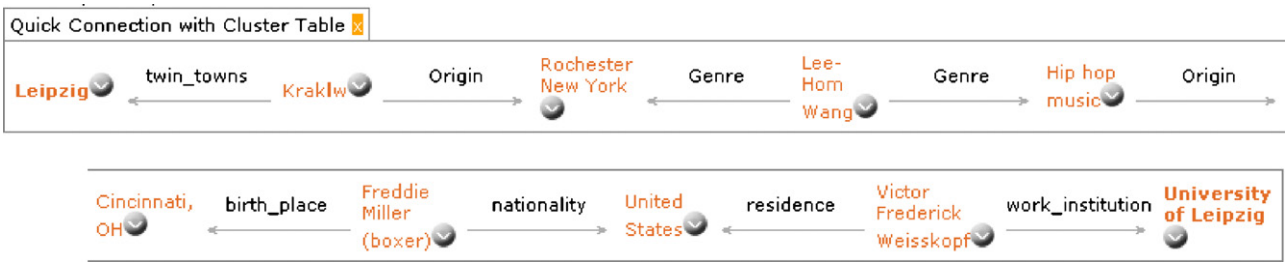

Fig. 9. The DBpedia Relationship Finder, displaying a connection between two objects.

submitting the query, the user is instantly informed whether a connection between the objects exists. If such a connection exists, the user can furthermore preview a connection between the objects, which is not necessarily the shortest (see Fig. 9). After that, several queries are posed to compute the shortest paths between the two objects, which are then displayed. Details of the used procedure can be found in [14].

6.3. Content annotation

Several applications have recently become publicly available that allow the annotation of Web content with DBpedia identifiers in an automated or semi-automated fashion. These annotations allow the discovery of Web content that is related to a DBpedia entity and enable third parties to enrich their content with data provided by DBpedia and interlinked data sources.

Muddy Boots[23] is a project commissioned by the BBC that aims to enhance the BBC news stories with external data. The Muddyboots APIs allow to identify the main actors (people and companies) in a BBC news story in an unambiguous way by means of DBpedia identifiers. In this way, the story is linked to DBpedia data, which is then used by a BBC prototype to populate a sidebar with background information on identified actors [12].
Open Calais[24] is a project by Thomson Reuters that provides a web service for named entity recognition from freetext as well as related tools. With the recent release 4 of the web service, entity descriptors are published as Linked Data with outgoing `owl:sameAs` links to DBpedia, Freebase and GeoNames. With this foundation, Thomson Reuters intends to publish several commercial datasets as Linked Data.
Faviki[25] is a social bookmarking tool that allows tagging of bookmarks with Wikipedia-based identifiers to prevent ambiguities. Identifiers are automatically suggested using the Zemanta API (see below). DBpedia is leveraged to view tags by topics and to provide tag descriptions in different languages.
Zemanta[26] provides tools for the semi-automated enrichment of blogs. The company offers its annotation engine to third parties via an API. Zemanta recently extended its API to generate RDF links pointing at DBpedia, Free-base, MusicBrainz and Semantic CrunchBase.
LODr[27] allows users to tag content that they contributed to popular Web 2.0 services (Flickr, del.icio.us, slideshare) using Linked Data identifiers, such as those provided by DBpedia. Tags may be translated semi-automatically into identifiers using Sindice.

Topbraid Composer[28] is a Semantic Web modeling environment that includes a built-in capability to resolve a label to a Wikipedia article, from which it derives a DBpedia resource URI. This functionality is also exposed to scripts using the SPARQLMotion scripting language.[29]

7. Related work

There is a vast body of works related to the semantification of Wikipedia. Comprehensive listings are provided by Michael Bergman[30] and by Wikipedia itself.[31] We will discuss some of the important approaches in the sequel.

7.1. Extraction of structured Wikipedia content

A second Wikipedia knowledge extraction effort is the Freebase Wikipedia Extraction (WEX) [15]. Free-base[32] is a commercial company that builds a huge online database which users can edit in a similar fashion as they edit Wikipedia articles today. Free-base employs Wikipedia knowledge as initial content for their database that will afterwards be edited by Freebase users. By synchronizing the DBpedia knowledge base with Wikipedia, DBpedia in contrast relies on the existing Wikipedia community to update content. Since November 2008, Freebase publishes its database as Linked Data and DBpedia as well as Freebase have set RDF links to same entities in the other data source.

A third project that extracts structured knowledge from Wikipedia is the YAGO project [19]. YAGO extracts 14 relationship types, such as `subClassOf`, `type`, `familyNameOf`, and `locatedIn` from different sources of information in Wikipedia. One source is the Wikipedia category system (for `subClassOf`, `locatedIn`, `diedInYear`, `bornInYear`), and another one are Wikipedia redirects. YAGO does not perform an infobox extraction like our approach. In order to improve the quality of its classification hierarchy, YAGO links leaf categories of the Wikipedia category hierarchy into the WordNet hierarchy. The YAGO and DBpedia projects cooperate and we serve the resulting YAGO classification together with the DBpedia knowledge base.

In [24] the KOG system is presented, which refines existing Wikipedia in-foboxes based on machine learning techniques using both SVMs and a more powerful joint-inference approach expressed in Markov Logic Networks. In conjunction with DBpedia, KOG could give Wikipedia authors valuable insights about inconsistencies and possible improvements of infobox data.

[23] http://muddyboots.rattleresearch.com/
[24] http://opencalais.com/
[25] http://faviki.com
[26] http://zemanta.com
[27] http://lodr.info/
[28] http://www.topbraidcomposer.com/
[29] http://www.topquadrant.com/sparqlmotion/smf.html#smf:dbpedia
[30] http://www.mkbergman.com/?p=417
[31] http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies
[32] http://www.freebase.com

Please cite this article in press as: C. Bizer, et al., DBpedia - A crystallization point for the Web of Data, Web Semantics: Sci. Serv. Agents World Wide Web (2009), doi:10.1016/j.websem.2009.07.002

7.2. NLP-based knowledge extraction

There is a vast number of approaches employing natural language processing techniques to obtain semantics from Wikipedia. Yahoo! Research Barcelona, for example, published a semantically annotated snapshot of Wikipedia,[33] which is used by Yahoo for entity ranking [25]. A commercial venture in this context is the Powerset search engine,[34] which uses NLP for both understanding queries in natural language as well retrieving relevant information from Wikipedia. Further potential for the DBpedia extraction as well as for the NLP-field in general lies in the idea to use huge bodies of background knowledge – like DBpedia – to improve the results of NLP-algorithms [11,8].

7.3. Stability of Wikipedia identifiers

Hepp et al. show in [10] that Wikipedia page IDs are reliable identifiers for conceptual entities and that they are stable enough to be used within knowledge management applications. Their finding confirms the approach of using DBpedia URIs for interlinking data sources across the Web of Data.

7.4. Advancing Wikipedia itself

The Semantic MediaWiki project [13,22] also aims at enabling the reuse of information within wikis as well as at enhancing search and browse facilities. Semantic MediaWiki is an extension of the MediaWiki software, which allows to add structured data into wikis using a specific syntax. Ultimately, the DBpedia and Semantic MediaWiki have similar goals. Both want to deliver the benefits of structured information in Wikipedia to the users, but use different approaches to achieve this aim. Semantic MediaWiki requires authors to deal with a new syntax and covering all structured information within Wikipedia would require converting all information into this syntax. DBpedia exploits the structure that already exists within Wikipedia. Therefore DBpedia's approach does not require changes from Wikipedia authors and can be employed against the complete content of Wikipedia. Both approaches could be combined synergetically by using DBpedia's extraction algorithms for existing Wikipedia content, while SMW's typed link functionality could be used to encode and represent additional semantics in wiki texts.

8. Conclusions and future work

The DBpedia project showed that a rich corpus of diverse knowledge can be obtained from the large scale collaboration of end-users, who are not even aware that they contribute to a structured knowledge base. The resulting DBpedia knowledge base covers a wide range of different domains and connects entities across these domains. The knowledge base represents the conceptual agreement of thousands of Wikipedia editors and evolves as conceptualizations change.

By allowing complex queries to be asked against Wikipedia content, the DBpedia knowledge base has the potential to revolutionize the access to Wikipedia. In the context of classic Web search engines, the knowledge base can be used to relate search terms to entities and to improve search results based on DBpedia's conceptual structure. The utility of the knowledge base as interlinking hub for the Web of Data is demonstrated by the increasing number of data sources that decide to set RDF links to DBpedia and the growing number of annotation tools that use DBpedia identifiers. Already today, the resulting Web of Data around DBpedia forms an exciting test-bed to develop, compare, and evaluate data integration, reasoning, and uncertainty management techniques, and to deploy operational Semantic Web applications.

As future work, the DBpedia project currently aims in the following directions:

Cross-language infobox knowledge fusion. Infoboxes within different Wikipedia editions cover different aspects of an entity at varying degrees of completeness. For instance, the Italian Wikipedia contains more knowledge about Italian cities and villages than the English one, while the German Wikipedia contains more structured information about persons than the English edition. By fusing infobox knowledge across editions and by applying different conflict resolution and consistency checking strategies within this process, it should be possible to derive an astonishingly detailed multi-domain knowledge base and to significantly increase the quality of this knowledge base compared to knowledge bases that are derived from single Wikipedia editions.

Wikipedia article augmentation. Interlinking DBpedia with other data sources makes it possible to develop a MediaWiki extension that augments Wikipedia articles with additional information as well as media items (pictures, audio) from these sources. For instance, a Wikipedia page about a geographic location like a city or monument can could augmented with additional pictures from Web data sources such as Flickr or with additional facts from statistical data sources such as Eurostat or the CIA Factbook.

Wikipedia consistency checking. The extraction of different Wikipedia editions and interlinking DBpedia with external Web knowledge sources lays the foundation for checking the consistency of Wikipedia content. For instance, whenever an Wikipedia author edits an infobox within a Wikipedia article, the content of the infobox could be checked against external data sources and the content of infoboxes within different language editions. Inconsistencies could be pointed out along with proposals on how to solve these inconsistencies. This way, DBpedia and the Web of Data could contribute back to the Wikipedia community and help to improve the overall quality of Wikipedia.

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: a nucleus for a web of open data, in: Proceedings of the 6th International Semantic Web Conference, 2007.
[2] S. Auer, J. Lehmann, What have innsbruck and leipzig in common? extracting semantics from wiki content, in: Proceedings of the 4th European Semantic Web Conference, 2007.
[3] C. Becker, C. Bizer, DBpedia mobile—a location-aware semantic web client, in: Proceedings of the Semantic Web Challenge, 2008.
[4] T. Berners-Lee, Linked Data-Design Issues, http://www.w3.org/DesignIssues/LinkedData.html (2006).
[5] C. Bizer, R. Cyganiak, T. Heath, How to publish Linked Data on the Web, http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/ (2007).
[6] C. Bizer, E. Pietriga, R. Lee, D. Karger, Fresnel: a browser-independent presentation vocabulary for rdf, in: Proceedings of the 5th International Semantic Web Conference, 2006.
[7] G. Cheng, W. Ge, H. Wu, Y. Qu, Searching semantic web objects based on class hierarchies, in: Proceedings of the 1st Linked Data on the Web Workshop, 2008.
[8] S. Cucerzan, Large-scale named entity disambiguation based on wikipedia data, in: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007.
[9] A. Harth, A. Hogan, J. Umbrich, S. Decker, Swse: objects before documents!, in: Proceedings of the Semantic Web Challenge, 2008.
[10] M. Hepp, K. Siorpaes, D. Bachlechner, Harvesting wiki consensus: using wikipedia entries as vocabulary for knowledge management, IEEE Internet Computing 11 (5) (2007) 54–65.
[11] J. Kazama, K. Torisawa, Exploiting wikipedia as external knowledge for named entity recognition, in: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007.
[12] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, R. Lee, C. Bizer, Media meets semantic web—how the bbc uses dbpedia and linked

[33] http://www.yr-bcn.es/dokuwiki/doku.php?id=semantically_annotated_snapshot_of_wikipedia
[34] http://www.powerset.com

Please cite this article in press as: C. Bizer, et al., DBpedia - A crystallization point for the Web of Data, Web Semantics: Sci. Serv. Agents World Wide Web (2009), doi:10.1016/j.websem.2009.07.002

data to make connections, in: Proceedings of the 6th European Semantic Web Conference, 2009.
[13] M. Krotzsch, D. Vrandecic, M. Volkel, Wikipedia and the semantic web—the missing links, in: Proceedings of Wikimania, 2005.
[14] J. Lehmann, J. Schiippel, S. Auer, Discovering unknown connections—the dbpedia relationship finder, in: Proceedings of the 1st SABRE Conference on Social Semantic Web, 2007.
[15] Metaweb Technologies, Freebase Wikipedia extraction (wex), http://download.freebase.com/wex/ (2009).
[16] F. Naumann, A. Bilke, J. Bleiholder, M. Weis, Data fusion in three steps: resolving schema, tuple, and value inconsistencies, IEEE Data Engineering Bulletin 29 (2) (2006) 21–31.
[17] Y. Raimond, C. Sutton, M. Sandler, Automatic interlinking of music datasets on the semantic web, in: Proceedings of the 1st Linked Data on the Web Workshop, 2008.
[18] A. Reka, B. Albert-Laszlo, Statistical mechanics of complex networks, Reviews of Modern Physics 74 (2002) 47–97.
[19] F.M. Suchanek, G. Kasneci, G. Weikum, Yago: a large ontology from wikipedia and wordnet, Journal of Web Semantics 6 (3) (2008) 203–217.
[20] T. Berners-Lee, et al., Tabulator: exploring and analyzing linked data on the semantic web, in: Proceedings of the 3rd International Semantic Web User Interaction Workshop, 2006.
[21] G. Tummarello, R. Delbru, E. Oren, Sindice.com: weaving the open linked data, in: Proceedings of the 6th International Semantic Web Conference, 2007.
[22] M. Volkel, M. Krotzsch, D. Vrandecic, H. Haller, R. Studer, Semantic wikipedia, in: 15th World Wide Web Conference, 2006.
[23] J. Volz, C. Bizer, M. Gaedke, G. Kobilarov, Silk—a link discovery framework for the web of data, in: Proceedings of the 2nd Linked Data on the Web Workshop, 2009.
[24] F. Wu, D. Weld, Automatically refining the wikipedia infobox ontology, in: Proceedings of the 17th World Wide Web Conference, 2008.
[25] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, G. Attardi, Ranking very many typed entities on wikipedia, in: 16th Conference on Conference on Information and Knowledge Management, 2007.

Please cite this article in press as: C. Bizer, et al., DBpedia - A crystallization point for the Web of Data, Web Semantics: Sci. Serv. Agents World Wide Web (2009), doi:10.1016/j.websem.2009.07.002