DBpedia as Entry Point in the Web of Data (Semantic Web)

Mile Gjorgjioski[1], Dijana Capeska-Bogatinoska[2] and Mirjana Trompeska[3]

Abstract – **Linked Data lies at the heart of what Semantic Web is all about: large scale integration of data on the Web. The importance of DBPedia is that it extracts structured information from Wikipedia and incorporates links to other datasets on the Web, e.g., to Geonames. By providing those extra links, in term of RDF triples, applications exploit the extra knowledge from other datasets when developing an application.**

Keywords – **semantic web, web of data, linked data, DBPedia.**

I. INTRODUCTION

Essentiality of the Semantic Web isn't just about putting data on the web, but in making links among them, so that a person or machine can explore the Web of data. With Linked Data, when you ask for some data, you can also find other related data, besides the one you were searching for. The goal of Linked Data is to enable people to share structured data on the Web, as easily as they can share documents today.

The term Linked Data, coined by Tim Berners-Lee, refers to a style of publishing interlinking structured data on the web. The value and usefulness of data increases the more if it is interlinked with other data. In summary, Linked Data is about using the web to create typed links between data from different sources.

So, to get Linked Data, we have to:

- Use the RDF data model to publish data on the web
- Use RDF links to interlink data from different data sources

Applying those principles, we can create a data community on the web, a space where people and organizations can post and consume data about anything. This community of data is called Web of data or Semantic Web.

The Web of data can be accessed using Linked Data browsers, such as Tabulator, Marbles, OpenLink RDF Browser or Disco, just as the traditional Web of documents is accessed using HTML browsers. Linked Data browsers enable users to navigate between different data sources by following RDF links, instead of following links between HTML pages.

The user can start with one data source, and then move through a potentially endless Web of data sources connected by RDF links. For example, a user can start with data about one person in one source, and it might be interested in information about person's home town. By following an RDF link, the user can navigate to that information contained in another dataset.

Just as the traditional document Web can be crawled by following hypertext links, the Web of data can be crawled by following RDF links. Search engines can provide sophisticated query capabilities, similar to those provided by conventional relational databases. Because the query results themselves are structured data, not just links to HTML pages, they can be immediately processed, thus enabling a new class of applications based on the Web of data.

A typical case of a large Linked Data is DBPedia, which makes the content of Wikipedia available in RDF. The importance of DBPedia is not only that it extract structured information from Wikipedia, but also that it incorporates links to other datasets on the Web, e.g., to Geonames. By providing those extra links, in term of RDF triples, applications may exploit the extra knowledge from other datasets when developing an application. DBPedia allows us to ask sophisticated queries against Wikipedia, and to link other datasets on the Web to Wikipedia data. This makes it easier for the huge amount of information in Wikipedia to be used in new and interesting ways, and also, this might inspire new mechanisms for navigating, linking and improving the encyclopedia itself.

II. RESOURCES AND RESOURCE IDENTIFIERS

If we want to publish data on the web, we first have to identify the item of interest in our domain. Resources can be anything: books, people, organizations, etc. They are things whose properties and relationships we want to describe. In web terminology, all items of interest are called *resources* [2].

Resources are named by Uniform Resource Identifiers (URIs), and this provides globally unique, distributed naming system that we need for distributed knowledge [3]. For anything that has URI, we say that is on the web. For example: you, the book which you have bought last week etc. URI have fundamental role in the Semantic Web; they keep the Web of data together. URI's work is not just as a name, but also as a mean of accessing information about a resource over the web.

III. RDF DATA MODEL AND RDF LINKS

RDF is an acronym for Resource Description Framework. It is a standard data and modeling specification used to encode metadata and digital information. The Semantic Web vision is

[1] Mile Gjorgjioski is with Omnia Computers, Pitu Guli No10, 7500 Prilep, Macedonia, E-mail: Mile.Gjorgjioski@omniacomputers.com

[2] Dijana Capeska-Bogatinosta is with Eurokompozit 11 Oktomvri AD, A.Makedonski 2/42, 7500 Prilep, Macedonia, E-mail: finance@eurokompozit.com.mk

[3] Mirjana Trompeska is with OU Kiril i Metodij, s.Kanatlarci, 7500 Prilep, Macedonia, E-mail: mirjanatrompeska@yahoo.com

predominantly based on the fundamental power of the RDF language. In RDF, a description of a resource is represented as a number of triples. The three parts of each triple are called: *subject, predicate* and *object*. For example:

Dijana	has the **email address**	**dijanac@excite.com**
(subject)	(predicate)	(object)

The subject of a triple is the URI identifying the described resource. The predicate indicates what kind of relation exists between subject and object; its URI too. The object can either be a simple literal value, or the URI of another resource that is somehow related to the subject.

Collection of RDF triples is called **RDF model**: *RDF triple is a labeled connection between two resources* [1]. **RDF links** represent links between two resources. An RDF link simply states that one piece of data has some kind of relationship to another piece of data. These relationships can have different types. For example, an RDF link that connects data about people can state that two people know each other. RDF links are the foundation for the Web of data.

IV. What is DBpedia

DBpedia is a community effort to extract structured information from Wikipedia and make them available on the Web. DBpedia allows us to ask sophisticated queries against Wikipedia and to link other data sets on the Web to Wikipedia data. DBpedia makes it easier for the huge amount of information in Wikipedia to be used in new and interesting ways [4].

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, and are created by relatively small groups of knowledge engineers, also, they 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 by making this information accessible on the Web.

The DBpedia knowledge base currently describes more than 3.4 million things, including 312,000 persons, 413,000 places, 94,000 music albums, 49,000 films, 15,000 video games, 140,000 organizations, 146,000 species and 4,600 diseases. The DBpedia data set features labels and abstracts for these 3.2 million things in up to 92 different languages. The DBpedia knowledge base altogether consists of over 1 billion pieces of information (RDF triples) out of which 257 million were extracted from the English edition of Wikipedia and 766 million were extracted from other language editions.

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 [4].

The DBpedia knowledge base allows you to ask quite surprising queries against Wikipedia, for instance "Give me all famous persons which were born in Skopje". The SPARQL query language is used to query this data.

The DBpedia knowledge base is served as Linked Data on the Web. DBpedia defines Linked Data URI for millions of concepts, and various data providers have started to set RDF links from their data sets to DBpedia, making DBpedia one of the central interlinking-hubs of the emerging Web of data. DBpedia uses RDF as a flexible data model for representing extracted information and for publishing it on the Web.

DBpedia uses RDF as flexible data model, for presenting the extracted data and their announcement on the web. Each "thing" in DBpedia set of data, is identified through URI reference in the format: http://dbpedia.org/resource/Name where "Name" is taken from the source article from Wikipedia, which has the following format:

http://en.wikipedia.org/wiki/Name. Each DBpedia resource is described with different characteristics. Resources are described with their label, short and long abstract on English language, link to the appropriate Wikipedia article and link to the picture that describes the "thing" (if it exists at all). If the "thing" has multilanguage version in Wikipedia, then a short and long abstracts on the appropriate language are added to the description, and links to the Wikipedia pages in that language.

DBpedia contains HTML links to outside web pages and RDF links to external sources of data. There are two kinds of HTML links: *dbpedia:reference* - which lead to couple of web pages for the "thing"; and *foaf:homepage* - links which lead to web pages which are considered to be official web pages (homepages) for the "thing".

For example, *dbpedia:reference* for Skopje is http://www.skopje.gov.mk/en, while the *foaf:homepage* is http://en.wikipedia.org/wiki/Skopje.

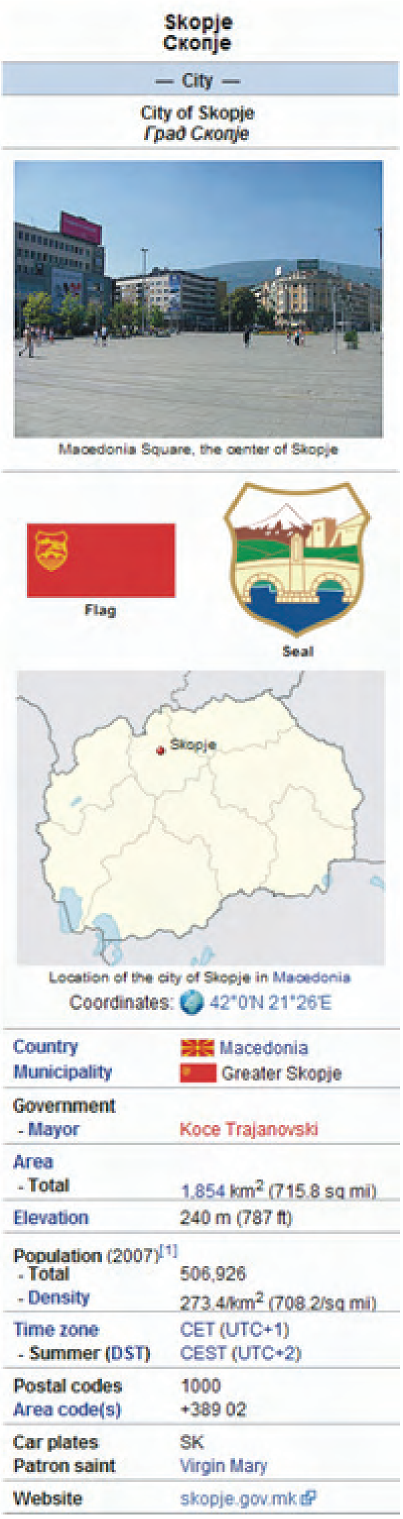
Wikipedia articles consist mostly of free text, but also contain different types of structured information, such as infobox templates, categorization information, images, geo-coordinates and links to external Web pages. For example, Fig. 1 illustrates structured infobox data from Wikipedia for the City of Skopje. Some of the extracted infobox information for the City of Skopje in DBpedia, are presented below:

```
@prefix dbpedia <http://dbpedia.org/resource/>
@prefix dbterm <http://dbpedia.org/property/>

dbpedia: Skopje
   dbterm:officialName Скопје
   dbterm:longd 21
   dbterm:longm 26
   dbterm:areaTotalKm 1854
   …….
   dbterm:LeaderName dbpedia:Koce Trajanovski
   ……..
   dbterm:location Skopje
   ……..
   dbterm:populationTotal 506926
   ……..
```

With DBpedia search engines, you can find all available information about the City of Skopje, in different languages. DBpedia data can automatically link with other data sets with

owl:sameAs, which mean that two or more resource are identical (it is the same resource, in this case – Skopje):
<http://dbpedia.org/resource/Skopje>
Owl:sameAs opencyc:Mx4rvVj-GZwpEbGdrcN5Y29ycA
Owl:sameAs http://sws.geonames.org/785842/
Owl:sameAs http://umbel.org/umbel/ne/wikipedia/Skopje



Skopje
Скопје

— City —

City of Skopje
Град Скопје

Macedonia Square, the center of Skopje

Flag

Seal

Location of the city of Skopje in Macedonia

Coordinates: 42°0′N 21°26′E

Country	Macedonia
Municipality	Greater Skopje
Government	
- Mayor	Koce Trajanovski
Area	
- Total	1,854 km^2 (715.8 sq mi)
Elevation	240 m (787 ft)
Population (2007)[1]	
- Total	506,926
- Density	273.4/km^2 (708.2/sq mi)
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)
Postal codes	1000
Area code(s)	+389 02
Car plates	SK
Patron saint	Virgin Mary
Website	skopje.gov.mk

Fig. 1. Structured information from Wikipedia

V. Access to DBpedia through the Web

DBpedia enables making queries which give us answers, based on the data from Wikipedia. It is made with the help of the query language SPARQL (http://dbpedia.org/sparql). As we have already said, Linked Data is method for announcement of RDF data on the web, and connecting the data between different sources of data. We can access them with Linked Data browser, same as the traditional web document is accessed with HTML browser.

So, instead following the links between HTML web pages, the Linked Data browser enables the user to move through different sources of data, while following the RDF links. This enables the user to start from certain source of data and after that to move through potentially endless web of sources, interweaved between them with RDF links.

This also helps the robots from the Linked Data browsers to follow these links, so they can collect data from the Semantic Web. DBpedia set of data, is represented as Linked Data. This gives us the opportunity to display DBpedia data with Linked Data browsers like: DISCO, Marbles, OpenLink Data Explorer, Tabulator or Zitgist Data Viewer.

In addition we have a few Linked Data URI, from the DBpedia set of data. In order to start surfing through the Semantic Web, enter some of these URI in some of the browsers named above:

- http://dbpedia.org/resource/The_Lord_of_the_Rings
- http://dbpedia.org/resource/Berlin
- http://dbpedia.org/resource/Semantic_Web

VI. SPARQL Query Interfaces

DBpedia provides a public SPARQL endpoint at http://dbpedia.org/sparql, which enables users to query the RDF data source with SPARQL queries such as:

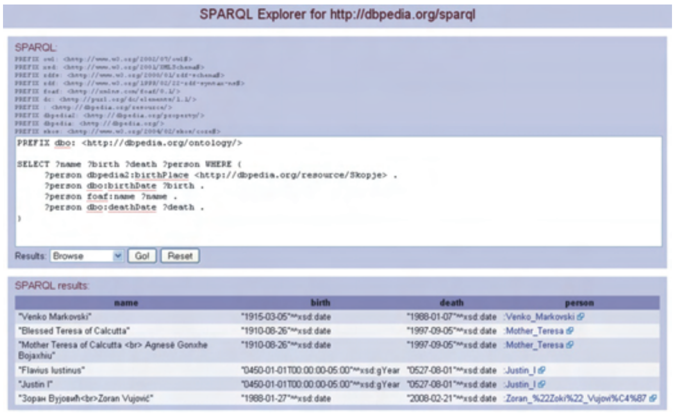
```
SELECT ?abstract
WHERE {
{ <http://dbpedia.org/resource/Skopje>
<http://dbpedia.org/property/abstract> ?abstract.
}
```

The query returns all the abstracts for the City of Skopje, given in each of the available languages.

The next query refines the abstracts returned to just the language specified, in this case ‘en’ (English):

```
SELECT ?abstract
WHERE {
{ <http://dbpedia.org/resource/Skopje>
<http://dbpedia.org/property/abstract> ?abstract.
FILTER langMatches (lang(?abstract), 'en')}
}
```

The SNORQL query explorer shown in the Fig. 2, provides a simpler interface to the DBpedia SPARQL endpoint. Fig. 2 shows both the query and the result returned for all famous persons who are born in Skopje.



```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?name ?birth ?death ?person WHERE {
     ?person dbpedia2:birthPlace <http://dbpedia.org/resource/Skopje> .
     ?person dbo:birthDate ?birth .
     ?person foaf:name ?name .
     ?person dbo:deathDate ?death .
}
```

name	birth	death	person
"Venko Markovski"	"1915-03-05"^^xsd:date	"1988-01-07"^^xsd:date	:Venko_Markovski
"Blessed Teresa of Calcutta"	"1910-08-26"^^xsd:date	"1997-09-05"^^xsd:date	:Mother_Teresa
"Mother Teresa of Calcutta Agnesë Gonxhe Bojaxhiu"	"1910-08-26"^^xsd:date	"1997-09-05"^^xsd:date	:Mother_Teresa
"Flavius Iustinus"	"0450-01-01T00:00:00-05:00"^^xsd:gYear	"0527-08-01"^^xsd:date	:Justin_I
"Justin I"	"0450-01-01T00:00:00-05:00"^^xsd:gYear	"0527-08-01"^^xsd:date	:Justin_I
"Зоран Вујовић Zoran Vujović"	"1988-01-27"^^xsd:date	"2008-02-21"^^xsd:date	:Zoran_%22Zoki%22_Vujovi%C4%87

Fig. 2. SNORQL query and result

VII. DBpedia Applications

This is the list of applications (in no particular order) to get us started using DBpedia [5]:

Faceted Browsers

- **Faceted Wikipedia Search** – allows us to explore Wikipedia via a faceted browsing interface;
- **OpenLink Virtuoso built-in Faceted Browser, and Search & Find Service**, on the DBpedia host instance – offers several paths of DBpedia data exploration, starting from Keyword or URI or Label

User Applications

- **DBpedia Mobile** – provides a map view annotated with DBpedia entities and information from other knowledge bases
- **DBpedia Relation finder** – enter objects and find out what connects them (Shockwave / Flash-based)
- **DBpedia Navigator** – navigate through DBpedia data

URI Lookup Services

- **DBpedia Lookup** – find DBpedia URI for keywords

Query Builders

- **OpenLink iSPARQL Visual Query Builder** – drag & drop visual interface for construction of SPARQL queries against DBpedia and/or other data sets
- **DBpedia query Builder** – build your own DBpedia queries

SPARQL query interfaces

- **SNORQL Query Builder** – query DBpedia using the SPARQL query language
- **OpenLink Virtuoso built-in SPARQL endpoint**, on the DBpedia host instance

Browser enhancements

- **DBpedia UserScript** – enhances Wikipedia pages with links to their corresponding DBpedia page

VIII. Conclusion

The Semantic Web represents evolutionary development of the World Wide Web, in which the meaning (semantics) of the information and the web services are defined, which helps the web to “understand” the user requirements. It appears as a result of the Tim Berners-Lee’s vision, as universal media for exchange of data, information and knowledge across the web. Tim Berners-Lee calls the resultant web of Linked Data, the Giant Global Graph, in contrast to the HTML-based World Wide Web.

The DBpedia knowledge base is served as Linked Data on the web. As DBpedia defines Linked Data URI for millions of concepts, various data providers have started to set RDF links from their data sets to DBpedia, making DBpedia one of the central interlinking-hubs of the emerging Web of data. Already today, the resulting Web of data around DBpedia, forms an exciting test-bed, to develop, compare and evaluate data integration, and to deploy operational Semantic Web applications.

References

[1] Ivan Herman, “Tutorial on the Semantic Web” – W3C: http://www.w3.org/People/Ivan/CorePresentations/SWTutorial/Slides.pdf , pp 42-45 and pp 131-140, 2010.

[2] Chris Bizer, Richard Cyganiak, Tom Heath, “How to Publish Linked Data on the Web” - http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/, 2008

[3] Joshua Tauberer, “What is RDF” – XML.com: http://www.xml.com/pub/a/2001/01/24/rdf.html?page=2

[4] DBpedia homepage, http://dbpedia.org/About

[5] DBpedia applications, http://wiki.dbpedia.org/Applications

[6] W3C Recommendation, “SPARQL Query Language for RDF” - http://www.w3.org/TR/rdf-sparql-query/ , 2008