Applying Linked Data Technologies for Online Newspapers

Tsvetanka Georgieva-Trifonova
Department of Mathematics and Informatics
University of Veliko Tarnovo
Veliko Tarnovo, Bulgaria

Tihomir Stefanov
Department of Mathematics and Informatics
University of Veliko Tarnovo
Veliko Tarnovo, Bulgaria

***Abstract*—The constantly growing data volume at the companies along with the necessity for finding information for the shortest possible time span involves methods of information search different from the ones conventionally used. The semantic technologies, developed in the late 90s and the beginning of the new century, are viewed as a new generation of databases and as text analyzing technologies. The present paper deals with researching the opportunities and examining the advantages of applying the semantic web technologies and the linked data for online newspapers. Besides, a RDF-based ontology for the purposes of a system for study and evaluation of online editions of regional daily newspapers is proposed. SPARQL endpoint is implemented to access the RDF data.**

Keywords—semantic web; ontology; linked data; SPARQL endpoint; RDF dataset; online newspaper

I. Introduction

The constantly growing data volume in companies along with the necessity for finding information within the minimum required time entails methods of information search different from conventional ones. The contemporary information society structure becomes more and more complex and the requirements to the effectiveness of the information processing algorithms are growing accordingly. The most popular technologies in this aspect recently are data mining, knowledge discovery in databases, machine learning. They provide theoretical and methodological basis for the study, analysis and rationalization of huge databases, but on account of the specifics of the Web data structure they themselves are not effective enough.

As a result, the semantic web popularity increases with its possibility to help in the dissemination of knowledge embedded in documents provided that the process of semantic interpretation becomes at least partially automated. Therefore, it is necessary to describe and present in sufficiently formalized form the data relevant to a specific area.

The present paper deals with researching the opportunities and introducing the advantages of applying the semantic web and linked data technologies for regional online newspapers in Bulgaria. The created RDF dataset *LinkedNewsData* which is linked to DBpedia and Europeana has been described. The data access is ensured by providing a SPARQL endpoint.

The rest of the paper is organized as follows. In Section 2, a review of the semantic web and linked data technologies is made. In Section 3, the related works on the application of semantic web technologies for digital newspaper archive maintenance are surveyed. In addition, the available linked data received from online newspapers along with SPARQL endpoint for data access are studied. In Section 4, the construction of the dataset *LinkedNewsData*, extracted from a system of study and evaluation of online regional newspaper editions, is motivated and described.

II. Semantic Web Technologies

The semantic technologies developed in the late 90s and the beginning of the new century, are viewed as a new generation databases and text analyzing technologies. The semantic web concept was proposed for the first time in 2001 by Tim Berners-Lee [1] – the founder of the World Wide Web. It consists in the automated process of conversion into semantic meaning of the data received from different network resources. The processing and exchange of information should be carried out not by people but by special agents (programs, distributed in the network). In order to be able to interact, these agents shall present the incoming data from each source in common form. Semantic web languages have two basic aspects. First, they shall have formal syntax and semantics to allow automated content processing. And second, a standard dictionary shall be provided, connected with the real world semantics, accessible for information sharing to both people and automated agents.

The Resource Description Framework (RDF) [2] standard, developed by the World Wide Web Consortium (W3C), allows for the semantic description of web resources and their relations in a way understandable to both people and machines, with the option XML presentation format (eXtensible Markup Language). RDF Schema is RDF extension, defining resource classes, their properties and their relations. OWL (Web Ontology Language) [3] is an extension of the RDF and RDFS, aimed specifically at description of reusable definitions of specific problem areas, called *ontologies*.

One of the main advantages of using RDF and OWL for the presentation of information is their reusability and their development through integration and upgrade of ontologies already built by other developers for specific areas, accessible on the Web. The development of semantic web languages is illustrated in Figure 1.

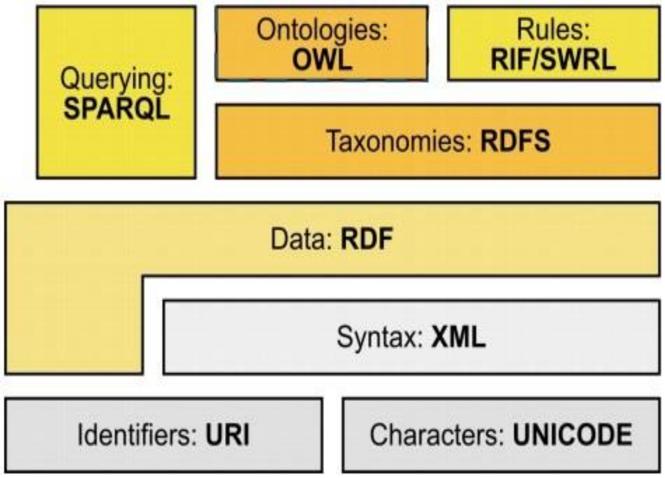


Fig. 1. Development of languages for semantic web

With the growing popularity of RDF and RDFS in the business companies and scientific circles, greater amount of RDF web content has been created which arises a necessity for standard access to the data stored. The language designed to allow the execution of queries to RDF data is SPARQL (SPARQL Protocol and RDF Query Language).

One of the most important concepts related to the semantic web, is the so-called *linked data*. Its goal is the publication of structured data, so that they can be easily linked to each other and thus, more useful. It is built on the basis of standard web technologies such as HTTP, URI, RDF, but develops them further so as to enable the sharing of information in a way comprehensible to computers. Some of the basic principles of linked data are:

- Using URI to designate the resources;
- Using HTTP URI for reference and search of resources by people and machines;
- The inclusion of links to related resources by using their URI when publishing a resource in the Web.

III. Applying the Semantic Web Technologies for Maintenance of the Digital Archive of a Newspaper

Usually, the technologies traditionally used provide opportunities for keyword search, for viewing the text of the published newspaper article, navigation through static hand-made hyperlinks between news materials (for example, links to previous topic related articles). Aspects that can be improved are:

- Search by keywords has limited expressive capability;
- Weak relationship between archive items: users may need to manually perform a few indirect requests before they can get answers to complex queries;
- The lack of common standard for the presentation by news archive sharing between the newspapers;
- Lack of consent about the terminology used in content description among journalists and between the journalists and archivists;
- Lack of commitment on the part of reporters in the archive creation process.

Neptuno project [4] is directed to the usage of the semantic web technologies to improve the exploitation and maintenance of the digital archive of a newspaper. The goal is to develop high-quality semantic archive for the Diari SEGRE newspaper, where:

- reporters and archivists have more expressive means to describe and annotate news items;
- reporters and readers are provided with better search and browsing options than those available;
- the archive maintenance system is open to integration in electronic marketplaces of news products.

As great investments have been made in the present news management systems, it is advisable to make this transition gradually. On account of this, in [5] the construction of an ontological framework based on existing journalistic and multimedia standards is proposed. These standards are based on XML technologies. Attached is the approach for the conversion of XML Schema into OWL, combined and supplemented with XML into RDF conversion. The main advantage of this approach is that it allows the reuse of existing metadata, which facilitates data integration, management and retrieval of previously stored news. The ontological framework is applied in the Diari Segre Media Group, which produces press, radio and television content.

A typical example illustrating the benefits of semantic technologies, is the BBC media group, which in its technological architecture replaces the MySQL with the semantic database OWLIM Enterprise of Ontotext, applied to the website, dedicated to the Football World Championship in 2010 and the Olympic Games in London 2012 [6]. Ontotext [7] is a Bulgarian company for semantic technologies, part of Sirma Group Holding.

Besides, BBC performs a successful integration and links data from different areas through applying the linked data technologies [8]. The linked data are accessible from the BBC SPARQL endpoint: http://api.talis.com/stores/bbc-backstage/services/sparql.

Some of the SPARQL endpoints providing access to linked data are:

- DBpedia endpoint: http://dbpedia.org/sparql;

DBpedia [9] is a dataset derived from Wikipedia, freely available through the use of the semantic web and the linked data technologies.

- Europeana endpoint: http://europeana.ontotext.com/sparql;

Europeana [10] is an electronic library in which scanned books, pictures, audio, video objects from museums and archives reflecting different aspects of European culture are being stored.

- British National Bibliography endpoint: http://bnb.data.bl.uk/sparql;

The British National Bibliography dataset [11] is designed to store information about the publishing of United Kingdom and the Republic of Ireland since 1950. The dataset includes metadata for published books, periodicals, magazines, newspapers, etc.

- Endpoints for science fiction datasets.

Dataset RKBExplorer [12] describes the publications, authors, institutions, conferences, etc.

- IEEE Papers (RKBExplorer) SPARQL Endpoint: http://ieee.rkbexplorer.com/sparql;
- DBLP Computer Science Bibliography (RKBExplorer) SPARQL Endpoint: http://dblp.rkbexplorer.com/sparql;
- ACM Bibliography (RKBExplorer) SPARQL Endpoint http://acm.rkbexplorer.com/sparql.

For the present paper, published linked data with a given SPARQL access endpoint have been studied. According to the study, there is a deficiency of datasets containing news from newspapers published in Bulgarian.

IV. *LinkedNewsData* – Linked Data Received from Online Newspapers

Except for the maintenance of the digital archive of the newspaper, the application of semantic web technologies is interesting in terms of constructing a system designed for the comparison and evaluation of online newspaper editions. The results generated from such a system, can be used from the owners of online editions in order that they be analyzed and thus become helpful in taking proper decisions for improvement of consumer access to information and satisfaction.

Below is shown a summary of the advantages of the semantic web data model to a database created for the purpose of such type of system.

- Without availability of semantic data, the owners of specific websites in which the search is based on SQL (Structured Query Language), are supposed to focus on the widely used data formats in order that the information shared is comprehensible to others. Deficiency is that in this way the sharing can not occur automatically and entails the human intervention.
- In terms of the data pattern in the semantic web, one and the same ontology, intended for the expression data meaning can be used by various websites that provide search through SPARQL.

The advantages mentioned above, are the basic reason to convert the existing relational database *SiteDB* into RDF data and their publishing as linked data.

The relational database *SiteDB* is designed and created for the purpose of a web-based system for examination and evaluation of online editions of regional daily papers, described in [13]. The database is implemented using the system for the management of relational databases MySQL.

In this paper, the *LinkedNewsData* ontology has been proposed, obtained after the conversion of the relational database *SiteDB* into RDF. *LinkedNewsData* is a set of RDF data associated with DBpedia and Europeana. Sample data for the study of the defined structure are accessible through the use of:

- SPARQL user applications – SPARQL endpoint: http://newspaper.byethost18.com/arc2-starter-pack/endpoint.php;
- Semantic web browsers – initial point: http://newspaper.byethost18.com/arc2-starter-pack/news.rdf.

The implementation of SPARQL endpoint is based on ARC RDF Store [14].

TABLE I. Types and Properties in *LinkedNewsData*

RDF type	**Property**	**Data types of the property values**	**Description**
Site	hasName	Text	Newspaper name
	owl:sameAs	URL	URL of the newspaper in DBpedia
	owl:sameAs	URL	URL of the newspaper in Europeana
	hasURL	URL	Website address of the newspaper
	whenUpdated	Datetime	Date and time of website latest update
	hasOutlinks	Integer	Number of external links referring to the URL of the Website
Page	ofSite	URL of sample type *Site*	Website of the first newspaper that published the news
	hasTitle	Text	Page title
	hasItemsCount	Integer	Number of news articles
	hasContent	Text	News content
	hasPageURL	URL	URL of the news article
	whenPageCreated	Datetime	Date and time of news article publishing
	hasExternalLinks	Integer	Number of external links
	hasWordsCount	Integer	Word count
	hasNewsImage	URL	URL of the news article picture
	inCategory	URL of sample type *Category*	Page category
Category	hasCategoryName	Text	Name of the category
	owl:sameAs	URL	URL of the category in DBpedia
	owl:sameAs	URL	URL of the category in Europeana
Word	hasWord	Text	A word
	owl:sameAs	URL	URL of the word in DBpedia
	owl:sameAs	URL	URL of the word in Europeana
SiteWord	inSite	URL of sample type *Site*	Website of the word containing newspaper
	containWord	URL of sample type *Word*	The word contained on the Website
	hasRepeatCount	Integer	Number of repetitions

Table 1 contains resume of the RDF types, defined in *LinkedNewsData*; the properties; the data type of the property values and their brief description. Graphic illustration of each type of samples (*Site*, *Page*, *Word*, *Category*, *SiteWord*) and some of their properties is shown in Figure 2.

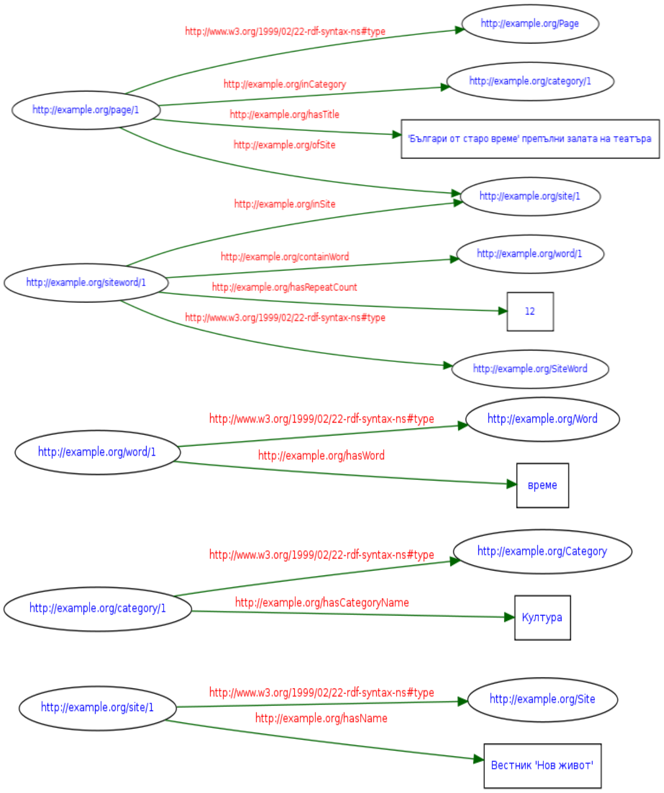

http://example.org/page/1
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://example.org/Page
http://example.org/inCategory
http://example.org/category/1
http://example.org/hasTitle
'Българи от старо време' препълни залата на театъра
http://example.org/ofSite
http://example.org/site/1
http://example.org/siteword/1
http://example.org/inSite
http://example.org/containWord
http://example.org/word/1
http://example.org/hasRepeatCount
12
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://example.org/SiteWord
http://example.org/word/1
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://example.org/Word
http://example.org/hasWord
време
http://example.org/category/1
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://example.org/Category
http://example.org/hasCategoryName
Култура
http://example.org/site/1
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://example.org/Site
http://example.org/hasName
Вестник 'Нов живот'

Fig. 2. RDF types *Site*, *Page*, *Word*, *Category*, *SiteWord* and some of their properties

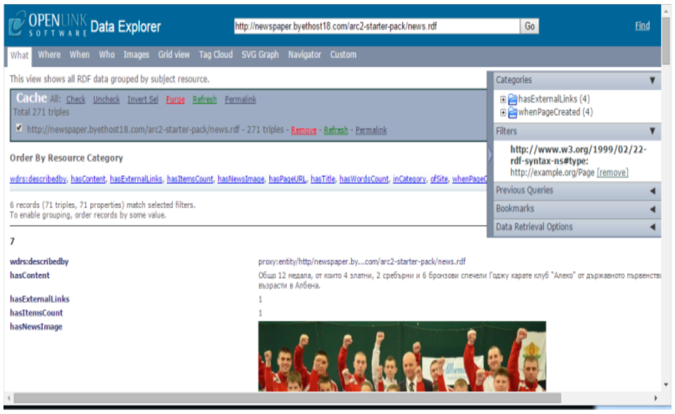
Figure 3 shows Turtle language presentation of the fragment from Figure 2.

```
@prefix e: <http://example.org/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .


<http://example.org/site/1> a e:Site ;
        e:hasName "Вестник \"Нов живот\"" .

<http://example.org/page/1> a e:Page ;
        e:ofSite <http://example.org/site/1> ;
        e:inCategory <http://example.org/category/1> ;
        e:hasTitle "\"Българи от старо време\" препълни залата на театъра" .

<http://example.org/category/1> a e:Category ;
        e:hasCategoryName "Култура" .

<http://example.org/siteword/1> a e:SiteWord ;
        e:inSite <http://example.org/site/1> ;
        e:containWord <http://example.org/word/7> ;
        e:hasRepeatCount "12"^^xsd:int .

<http://example.org/word/1> a e:Word ;
        e:hasWord "време" .
```

Fig. 3. Turtle language presentation

Figure 4 shows a view of the DataSet *LinkedNewsData* in OpenLink Data Explorer, which is a browser extension for semantic data browsing.

Some general data about RDF dataset *LinkedNewsData* could be collected through the execution of SPARQL queries, as shown in Table 2.



This view shows all RDF data grouped by subject resource.

Total 271 triples

☑ http://newspaper.byethost18.com/arc2-starter-pack/news.rdf - 271 triples - Remove - Refresh - Permalink

http://www.w3.org/1999/02/22-rdf-syntax-ns#type:
http://example.org/Page [remove]

Order By Resource Category

wdrs:describedby, hasContent, hasExternalLinks, hasItemsCount, hasNewsImage, hasPageURL, hasTitle, hasWordsCount, inCategory, ofSite, whenPageC

6 records (71 triples, 71 properties) match selected filters.
To enable grouping, order records by some value.

7

wdrs:describedby	proxy:entity/http/newspaper.by...com/arc2-starter-pack/news.rdf
hasContent	Общо 12 медала, от които 4 златни, 2 сребърни и 6 бронзови спечели Годжу карате клуб "Алеко" от държавното първенств възрасти в Албена.
hasExternalLinks	1
hasItemsCount	1
hasNewsImage	

Fig. 4. View of *LinkedNewsData* in OpenLink Data Explorer

TABLE II. SPARQL QUERIES FOR RETRIEVING GENERAL DATA ABOUT *LINKEDNEWSDATA*

Number of triplets	SELECT (COUNT(*) AS ?c)
	WHERE {?s ?p ?o.}

Number of newspapers	PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
	PREFIX np:<http://example.org/>
	
	SELECT (COUNT (?s) AS ?c)
	WHERE { ?s rdf:type np:Site. }

Number of pages per newspaper	PREFIX np:<http://example.org/>
	
	SELECT ?sn (COUNT (?p) AS ?c)
	WHERE { ?p np:ofSite ?s.
	?s np:hasName ?sn. }
	GROUP BY ?sn

Number of pages per category	PREFIX np:<http://example.org/>
	
	SELECT ?cn (COUNT(?p) AS ?cp)
	WHERE { ?p np:inCategory ?c.
	?c np:hasCategoryName ?cn. }
	GROUP BY ?cn

Number of words per newspaper	PREFIX np:<http://example.org/>
	
	SELECT ?sn (COUNT (?sw) AS ?c)
	WHERE { ?sw np:inSite ?s.
	?s np:hasName ?sn. }
	GROUP BY ?sn

Total number of the repetitions of a word in all newspapers	PREFIX np:<http://example.org/>
	
	SELECT ?wn (SUM(?rc) AS ?sumrc)
	WHERE { ?sw np:inSite ?s;
	np:containWord ?w;
	np:hasRepeatCount ?rc.
	?w np:hasWord ?wn.
	}
	GROUP BY ?wn
	ORDER BY DESC(?sumrc)

Number of links to DBpedia, of links to Europeana	SELECT (COUNT(?l) AS ?c)
	WHERE { ?s owl:sameAs ?l. }

The linking of dataset *LinkedNewsData* with DBpedia and Europeana is carried out by means of Google refine. The connection is made through the names of the newspapers, the news categories and the words contained in online newspaper websites.

V. Conclusion

In this paper, the opportunities and the advantages of semantic web technologies usage are discussed in terms of maintenance of a digital archive of a newspaper.

The transformation of an existing relational database that stores data for the online editions of regional newspapers into RDF format and their publishing as linked data are dealt with. Our future work envisages creation of user friendly interface to access the data through a simple browser.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web, Scientific American”, 2001, pp. 35-43, available at: http://www.w3.org/2001/sw (accessed 5 May 2015).

[2] D. Brickley and R.V. Guha, “RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation”, 2004, available at: http://www.w3.org/TR/rdf-schema (accessed 5 May 2015).

[3] D. L. McGuinness and F. Harmelen, “OWL Web Ontology Language Overview, W3C Recommendation”, 2004, available at: http://www.w3.org/TR/owl-features (accessed 5 May 2015).

[4] P. Castells, F. Perdrix, E. Pulido, M. Rico, R. Benjamins, J. Contreras, and J. Lorés, “Neptuno: semantic web technologies for a digital newspaper archive”, The Semantic Web: Research and Applications: First European Semantic Web Symposium, Berlin, Springer, 2004, pp. 445-458.

[5] R. García, F. Perdrix, R. Gil, and M. Oliva, “The Semantic Web as a Newspaper Media Convergence Facilitator”, Journal of Web Semantics, Vol. 6, No. 2, 2008, pp. 151-161.

[6] TechNews.bg, “BBC uses Semantic Technologies of the Bulgarian Company Ontotext”, 2010, available at: http://technews.bg/article-18510.html#.U5RGPfl_utM (accessed 5 May 2015).

[7] Ontotext, “Ontotext provides a complete set of semantic technologies including text mining and GraphDB™, an RDF triplestore that performs inferencing at scale”, 2011, available at: http://www.ontotext.com/company (accessed 5 May 2015).

[8] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee, “Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections”, The Semantic Web: Research and Applications, Lecture Notes in Computer Science, Spinger, Vol. 5554, 2009, pp. 723–737.

[9] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Kleef, S. Auer, and C. Bizer, “DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia”, Semantic Web – Interoperability, Usability, Applicability, IOS Press, 2012, pp. 1-28.

[10] N. Ikonomov, B. Simeonov, J. Parvanova, and V. Alexiev, “Europeana Creative. EDM Endpoint. Custom Views.”, Digital Presentation and Preservation of Cultural and Scientific Heritage, Vol. 3, No 1, 2013, pp. 35-43.

[11] C. Deliot, “Publishing the British National Bibliography as Linked Open Data”, Catalogue & Index, Issue 174, 2014, pp. 13-18.

[12] H. Glaser, I. Millard, and A. Jaffri, “Rkbexplorer.com: A knowledge driven infrastructure for linked data providers”, In European Semantic Web Conference, 2008, pp. 797–801.

[13] T. Stefanov and D. Tsvetkov, “A Model for Evaluation of Regional Electronic Media in terms of Efficiency Criteria and User Satisfaction”, Collection of Writings ‘Days of Science 2014’, Veliko Turnovo, 2014, in press.

[14] A. McIntyre and E. Durham, “ARC RDF Store Easy RDF and SPARQL for LAMP systems”, CSC 8711 Project 4, 2011.