INCORPORATING A SEMANTICALLY ENRICHED NAVIGATION LAYER ONTO AN RDF METADATABASE

TERESA SUSANA MENDES PEREIRA; ANA ALICE BAPTISTA

Universidade do Minho
Campus de Azurém, 4800-058, Guimarães, Portugal
tpereira@dsi.uminho.pt
analice@dsi.uminho.pt

Information Society Technologies (IST) funded Omnipaper project, proposes to investigate efficient ways to enable an access to distributed, and heterogeneous digital news archives through the use of state-of-the-art technologies such as RDF, and XTM. In the Omnipaper project we intend to achieve the implementation of a final prototype that enables users (professional journalists and occasional users) to have simultaneous and structured access to the articles of a large number of digital European news providers. This paper proposes to describe the work developed in the Omnipaper RDF prototype focusing the use of the IPTC Subject Codes in order to incorporate a semantically enriched navigation layer onto an RDF/XML metadata descriptions developed in the RDF prototype.

Keywords: metadata; Resource Description Framework (RDF); IPTC subject codes; ontology; navigation.

1. INTRODUCTION

During the last decades, the amount of digital information has grown exponentially. More and more information is becoming available in electronic form and its accessibility is, in terms of network presence, increasing rapidly. With this growth in availability, the need of information coupling has grown as well. Since it is physically becoming easier to compare information from geographically spread sources, the need for coupling information on a semantic level is on the rise [1]. The important challenge of the Web researchers resides in the need of organizing the immeasurable number of Web pages that appear everyday at every hour in the Internet.

The IST-funded OmniPaper project (Smart Access to European Newspapers, IST-2001-32174) investigates ways for drastically enhancing access to many different types of distributed information resources. One of the principal aspects of this project is the whole metadata layer, purposed to describe the content of news articles from distributed and heterogeneous resources, in order to provide a more efficient resource discovery on the Web. Conceptually the Omnipaper architecture is presented with two metadata layers, the Local Knowledge Layer and the Overall Knowledge Layer. Both layers were implemented in simultaneously using two different metadata approaches: the Resource Description Framework (RDF) approach, and the Topic Maps (TM) approach.

The main purpose of the Local Knowledge Layer is to provide a standard semantic description of all the existent articles in order to enable a structured and uniform access to the available distributed archives, while the Overall Knowledge Layer is a higher abstraction level that provide a common access user interface by integrating and relating the metadata information coming from the local knowledge layer. This layer intends to provide a cross-archive linking and a multi-archive navigation through the metadata news information.

In the scope of the Omnipaper project, it was implemented in simultaneously two different and independent prototypes with the same functionalities, through the use of state-of-the-art technologies: the RDF technology and the Topic Maps technology.

This paper intends to present, on particularly of RDF approach, the development of the search functionalities, defined on the semantically and syntactically common set of metadata elements, with the navigation and browsing functionalities in the IPTC Subject Codes tree, providing a resource discovery on the Web easier and more efficient, according to the user requirements.

This paper explains the work conducted in the representation of the IPTC Subject Codes, implemented in the scope of the Omnipaper RDF prototype, and is structured as follows: in Section 2 the term Ontology will be presented and some ontology's representation languages will be analysed. In section 3 the work conducted in the implementation of the IPTC Subject Codes Ontology will be presented . Lastly, the Section 4 it the conclusions and the future work will be presented.

2. ONTOLOGY

Traditionally the term '*Ontology'* is a branch of metaphysics to provide a definitive and exhaustive classification of the nature of human beings [2]. Now it is commonly used in several domains in particular in the information science area, to support the sharing and reuse of the formally represented knowledge, contributing to the definition of the common vocabulary in which shared knowledge is represented [3]. Furthermore, in logic programming area, the ontologies are defined with two main functions: (1) "Provide a way of viewing the world, and hence for organising information"; (2) "The ontologies are required for interoperability, to define a shared vocabulary and meanings for terms with respect to other terms" [3]. The ontologies are organized and structured by concepts, not by words [4].

An ontology language provides semantics for a set of concepts and relations in order to produce the qualified and the possible interpretations. Some languages allow the definition of axioms or logical relations between terms for the same purpose [5]. The selection of ontology language is based on the type of knowledge structures we intend to represent. Therefore, in the scope of the Omnipaper project, particularly on the RDF approach, a deep studied of several ontology language that can best fit our description needs was developed, especially to achieve the description purpose of the hierarchical tree represented in the IPCT Subject Codes. In the following sub-sections some initiatives studied in representing ontologies will be presented. This selection is based on languages associated with the Semantic Web.

2.1 RDF-S

The RDF-S is the first to be presented because of its acceptance in the Semantic Web community, and since then it was our selection the in the representation of the IPTC Subject Codes structure. The RDF-S [6] is an official recommendation of the World Wide Web (W3C) [7]. RDF is a knowledge representation format intended to describe metadata information of resources on the World Wide Web. RDF Schema specification refers to itself as the RDF Vocabulary Description Language [8]. In particular, the RDF-S defines classes and properties that might be used to describe classes, properties and other resources. It provides mechanisms for describing groups of related resources and the relationships between resources. These resources are used to determine characteristics of other resources, such as the domains and ranges of properties [8].

2.2. OIL

There is a number of possible meanings of the acronym: "Ontology Inference Layer", or "Ontology Interchange Language", but all of them contain the word "Ontology" [9]. The OIL's presentation syntax is intended for human readers and writers of OIL ontologies. For machines, OIL uses RDF as its syntax. OIL exploits as much as possible the modelling primitives of RDF-S. This provides crucial backwards compatibility, allowing OIL ontologies to be treated as extensions of RDF and RDF-S ontologies, and making OIL ontologies available not only to OIL-aware applications, but also to applications that are only RDF-aware [9].

2.3. DAML

The DARPA Agent Markup Language (DAML) program officially began in August 2000. The goal of the DAML effort is to develop a language and tools to facilitate the concept of the Semantic Web [10]. In August 2000, it was released DAML-ONT, a simple language for expressing more sophisticated RDF class definitions than permitted by RDF-S. DAML is an extension of the RDF-S adding up more semantics to the data [11]. The DAML group soon pooled efforts with the Ontology Inference Layer (OIL). The result of these efforts is DAML+OIL, a language for expressing far more sophisticated classifications and properties of resources than RDF-S [10].

Figure 1 illustrates the DAML+OIL ontology language as RDF-S extension:

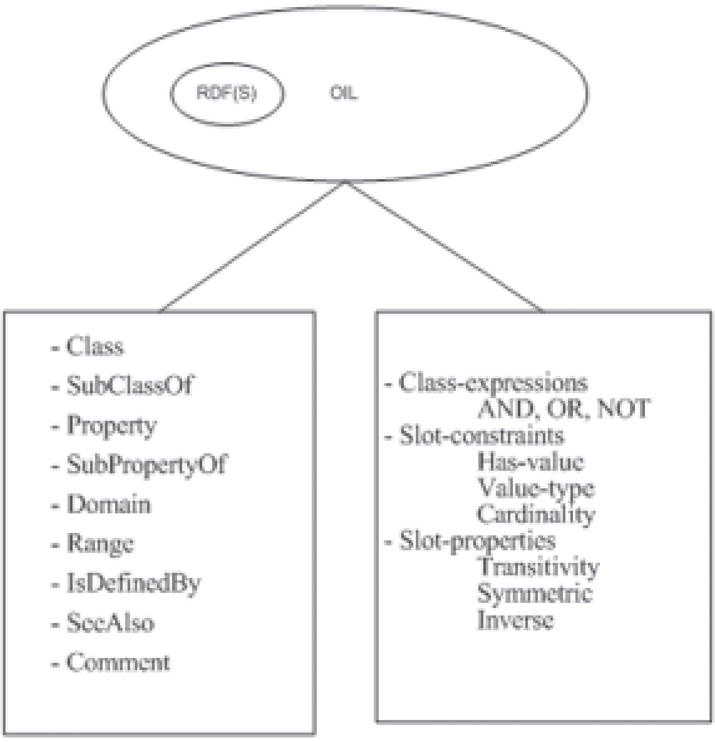

FIGURE 1. DAML+OIL AS RDF-S EXTENSION, ADAPTED FROM [12]

2.3. OWL

The Web Ontology Language OWL is a semantic markup language for publishing and sharing ontologies on the World Wide Web. OWL is developed as a vocabulary extension of RDF and is derived from the DAML+OIL Web Ontology Language [9, 13]. The OWL Web

Ontology Language is intended to provide a language that can be used to describe the classes and relations between them that are inherent in Web documents and applications [14]. The usually uses of the OWL language are:

1. Formalize a domain by defining classes and properties of those classes;

2. Define individuals and assert properties about them.

The OWL ontology is represented as a set of RDF triples. As with any set of RDF triples, OWL triples can be represented in many different syntactic forms [15].

OWL is a vocabulary extension of RDF. Thus any RDF graph forms the OWL ontology. Further, the meaning given to an RDF graph by OWL includes the meaning given to the graph by RDF. OWL ontology's can thus include arbitrary RDF content, which is treated in a manner consistent with its treatment by RDF. OWL assigns an additional meaning to certain RDF triples. The OWL Abstract Syntax and Semantics specifies exactly which triples are assigned a specific meaning, and what this meaning is [9].

Considering the main features of the expressive Semantic Web languages presented above, and due to the simple hierarchy of the concepts and relations introduced in the IPTC Subject Codes structure, the RDF-S was our ontology language selection to represent the hierarchical tree of the IPTC Subject Codes included in the implementation of the RDF prototype and developed in the scope of the Omnipaper project. This choice was based on the fact that the IPTC Subject Codes structure is so simple that the use of a more expressive language wouldn't accomplish any beneficial use.

3. IMPLEMENTATION

The research work conducted in the implementation of the Omnipaper RDF approach proceeds to accomplish the purpose on (1) definition of an application profile with all the necessary metadata elements, selected to describe the news articles [16]; (2) definition of a metadatabase that contains the metadata information of the news articles [17]; (3) definition of an ontology to represent the hierarchical concepts of the IPTC Subject Codes. The development of these issues leaded to the implementation of two RDF prototypes. In the first one the two first steps were performed. The second one is developed in order to add value to the first Omnipaper RDF prototype, which includes the third step.

In the implementation of the first RDF prototype, developed in the scope of the Omnipaper project, the digital news articles were catalogued according to standard and normalized metadata vocabularies [17], in order to enhance the search facility, through the RDF/XML metadata description stored in a metadatabase, enabling a structured and uniform access to the available distributed archives, which are the providers of the digital news. This first RDF prototype was developed in the Local Knowledge Layer.

In the Overall Knowledge Layer the second RDF prototype was implemented, including the navigation functionality through the hierarchical tree of the IPTC Subject Codes, enabling browsing through the organized concepts represented on the IPTC Subject Codes.

The IPTC Subject Codes is constituted with a hierarchical three-level tree of subject codes, which describes the content of a set of terms. Topics of Subject level provide a description of the editorial content of news, a SubjectMatter provides a description at a more precise level and finally a SubjectDetail at a more specific level. To represent the International Press Telecommunications Council Subject Codes, several ontology languages were analysed and studied in order to select the one that best fits its hierarchical representation. However the IPTC Subject

Codes in the semantically point of view, are not that rich, and due to its simplicity, it was only necessary to define its hierarchical concepts. Therefore as it was stated above, the RDF-S was the sufficient language to complete the representation of the hierarchical tree represented in the IPTC Subject Codes.

After the implementation of the IPTC Subject Codes through the RDF-S representation language, it was included in a metadatabase. The connection with the subject elements included in the hierarchical tree of the IPTC Subject Codes is made through the metadata element *"dc:subject"*. Furthermore, in the application profile definition, the "*rds:range*" of the metadata element "*dc:subject*" are the IPTC Subject Codes.

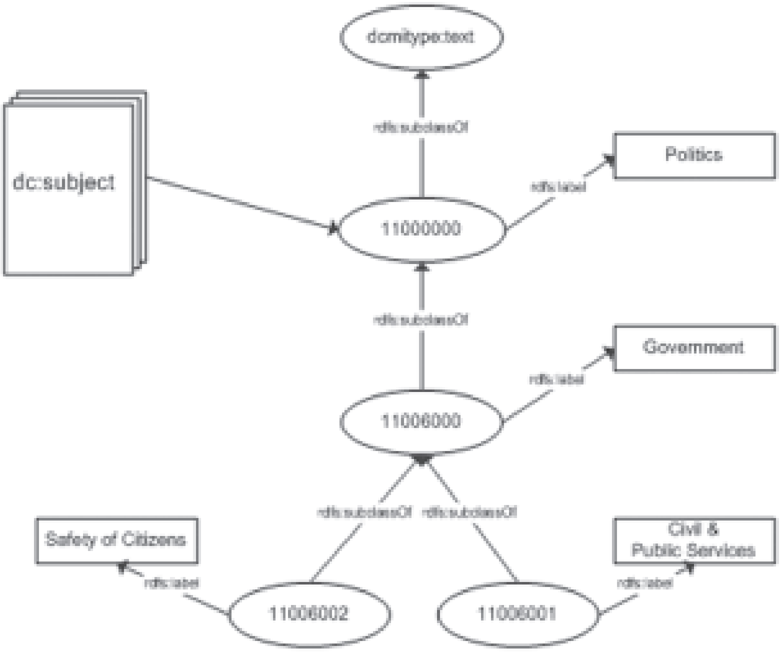

FIGURE 2. EXAMPLE OF PART OF THE 'POLITICS' SUBJECT BRANCH MODELLED IN RDF-S

The RDF prototype was developed over RDF Gateway [18], since RDF Gateway can generate HTML pages to users, and can also create a Native Database and a Package that are by themselves the applications. RDF Gateway is both a Web client and a Web server with a native RDF database for managing information on a server. The data access is made via HTTP.

4. CONCLUSION

This paper introduces the representation of a very small ontology, defined through the use of a state-of-art semantic technology, such as RDF-S. A set of the expressive Semantic Web languages was covered and studied, proceeding to the selection of one that best fits our description needs. As it was stated above, we conclude that the RDF-S completes all our needs in the description of the hierarchical tree of the IPTC Subject Codes. In particular, the notion of the ontology term was looked at and since then several interpretation of this term were found, and the one that better serve our purpose was presented.

In order to add value to the first prototype developed in the Omnipaper RDF prototype the navigation functionality was added, enabling the search of semantic concepts in the set of subject elements defined in the hierarchical subject tree performed in the IPTC Subject Codes.

The Omnipaper RDF prototype will be one of the first studies on how metadata and RDF contribute to information retrieval heterogeneous digital resources when compared with other approaches developed in the Omnipaper project. Furthermore, the performance of the IPTC Subject Codes ontology makes the prototype much more powerful in navigation of information resources. In the next phase the automatic link extraction will be implemented, proceeding to the cross-archive purposed to automatically extract links between a given news article and related articles and then store this information in the link database. Actually, we are currently working on the multilingual process.

REFERENCES

1 Lassila, O.R.S., Ralph, ***Resource Description Framework (RDF) Model and Syntax Specification***. 1999, from http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

2 *NetLing - Dictionary*, from http://www.linktotal.net/tp.htm?http://www.onelook.com/

3 Miller, L., *Ontologies and Metadata*, from http://ilrt.org/discovery/2000/11/lux/

4 *Ontology (computer science)*, from http://en.wikipedia.org/wiki/Ontology_(computer_science)

5 Mika, P., *Applied Ontology-based Knowledge Management: A Report on the State-of-the-Art*, in *Master*. 2002, Vrije Universiteit: Amsterdam, from http://www.cs.vu.nl/~pmika/thesis/pmika-thesis-full.doc

6 *RDF Vocabulary Description Language 1.0: RDF Schema*. 10 Febuary 2004, from http://www.w3.org/TR/rdf-schema/

7 *W3C World Wide Web Consortium*, from http://www.w3c.org/

8 Guha, D.B.R.V., *RDF Vocabulary Description Language 1.0: RDF Schema*, from http://www.w3.org/TR/2002/WD-rdf-schema-20021112/

9 Horrocks, F.v.H.I., *Questions and answers on OIL: the Ontology Inference Layer for the Semantic Web*, from http://www.ontoknowledge.org/oil/oil-faq.html

10 Ouellet, U.O.R., *DAML Reference*. 01 May 2002, from http://www.xml.com/lpt/a/2002/05/01/damlref.html

11 Garshol, L.M., *Topic maps, RDF, DAML, OIL - A comparison*, from http://www.ontopia.net/topicmaps/materials/tmrdfoildaml.htm

12 Stuckenschmidt, H., *DAML+OIL Overview*.

13 Horrocks;, F.v.H.J.H.I., *Web Ontology Language (OWL)Reference Version 1.0*, from http://www.w3.org/TR/2003/WD-owl-ref-20030221/

14 McGuinness, M.K.S.D.L., *Web Ontology Language (OWL) Guide Version 1.0*. 4 November 2002, from http://www.w3.org/TR/2002/WD-owl-guide-20021104/

15 Beckett, D., *Coments to Journal archives.* Journalblog, 2003.

16 Teresa Pereira, A.A.B., Tomoko Yaginuma. *Perfil de Aplicação e Esquema RDF dos Elementos de Metadados do Projecto Omnipaper.* in *CLME'2003 -3ºCongresso Luso-Moçambicano de Engenharia*. 2003. Maputo, Moçambique.

17 Tomoko Yaginuma, T.P., Ana Alice Baptista. *Metadata Elements for Digital News Articles in the Omnipaper Project*. in *ELPUB 2003 - 7th ICCC/IFIP International Conference on Electronic Publishing*. June 2003. University of Minho - Guimarães, Portugal.

18 *RDF Gateway A platafform for the semantic Web,* from http://www.intellidimension.com/