THE EXTENSION OF THE OMNIPAPER SYSTEM IN THE CONTEXT OF SCIENTIFIC PUBLICATIONS

Teresa Susana Mendes Pereira[1], **Ana Alice Baptista**[2]

Abstract Today the Internet is an important information source, which facilitates the search and access to information contents on the Web. In fact, the Internet has become an important tool used daily by scholars in the development of their work. However the contents published on the Web increase daily and consequently difficult the identification of new contents published in various information sources. In this context the RSS technology introduces a new dimension in the access and distribution mechanisms of new contents published by distributed information sources. In the scope of scientific contents the use of RSS technology helps the scholars to be up to date of new scientific resources provided by several and distributed information sources. An instance of the OmniPaper RDF prototype has been developed in order to instantiate the mechanisms of distributed information retrieval investigated in the context of the news published in newspapers and use them in the context of scientific contents. In addition a central metadatabase was developed through the RSS approach, in order to enable the scientific content syndication. This paper intends to describe the steps involved in the development of the instantiation system of the OmniPaper RDF prototype.

[1] Teresa Susana Mendes Pereira

Polytechnic Institute of Viana do Castelo, Superior School of Business Studies, Valença, Portugal, e-mail: tpereira@esce.ipvc.pt

[2] Ana Alice Baptista

University of Minho, School of Engineering, Information System Department, Guimarães, Portugal, e-mail: analice@dsi.uminho.pt

1 Introduction

Today the Web is an important and widely used information source. In fact the increased use of the Web associated to the constant evolution of technologies has promoted the development of sophisticated information systems to facilitate the access and dissemination of scientific contents produced by scientific communities. However, the Web provides several information sources and, consequently, the identification of new contents or updates demands time. In fact, users spend a lot of time tracking a set of information sources to check for new contents or updates and sometimes some of the resources are not even accessed.

RSS is an XML-based format to syndicate information, or metadata. The content syndication helps the user to be up to date of new contents published in different and distributed information sources and improves the visibility of the contents published, guaranteeing that the user becomes aware when new contents are published.

In the context of scientific contents the use of the RSS technology introduces important advantages in the distribution and dissemination process of the results produced by scientific communities. In spite of the RSS being primarily used in relaying the latest entries' headlines of the newspapers and weblogs, it has been adapted to a wide range of uses in the description of Web contents, to enhance the rapid dissemination of the contents. Some journals already use the RSS format in the description of the scientific articles published, such as D-Lib, Ariadne, Nature Publishing Group's (NPG), et cetera. In fact the syndication of metadata information of scientific contents has been contributing to the transformation of the current communication processes and information retrieval systems.

In the OmniPaper project were investigated and developed sophisticated mechanisms of distributed information retrieval in order to facilitate the access and distribution of the news contents published in the newspapers. These functionalities were supported by a semantic metadata layer. In order to take advantage of the search and browsing functionalities developed in the OmniPaper project, an instance of the OmniPaper RDF prototype was created in the context of scientific publication. In fact the whole RDF structure (described below) developed in the OmniPaper was used in the context of scientific contents in order to provide a mechanism to facilitate the distribution and dissemination of scientific research developments. The metadata layer followed an RSS approach in order to enable the syndication of the metadata information of the scientific contents.

This paper intends to describe the research work conducted in the implementation of the instantiated system of the OmniPaper RDF prototype and deepens each step shown below.

2 The RDF Prototype Developed in the OmniPaper Project

The OmniPaper (Smart Access to European Newspapers, IST-2001-32174) was a project from the European Commission IST (Information Society Technologies) program. The OmniPaper project aims to (1) find and test mechanisms for retrieving information from distributed sources in an efficient way; (2) Find and test ways for creating a uniform access point to several distributed information sources; (3) Make this access point as usable and user-friendly as possible; (4) Lift widely distributed digital collections to a higher level.

One of the principal aspects of the project is the whole metadata layer of the system that describes the metadata information of each local archive of digital news providers. This approach enables the user to search and navigate on the metadata information instead of performing integral text search, as it is usual in the most general information retrieval system. The OmniPaper architecture has two metadata layers: the Local Knowledge Layer and the Overall Knowledge Layer. The Local Knowledge Layer is composed of distributed metadatabases that contain standard semantic descriptions of all the existent articles provided by each digital news provider; this, enables a structured and uniform access to the available distributed archives. The Overall Knowledge Layer includes a conceptual layer which effort is to facilitate cross-archive browsing and navigation through the use of a web of concepts. Furthermore, this layer is enriched with multilingual and personalization functionalities for the interface with the user [15].

The steps followed in the development of the OmniPaper RDF prototype were [1]:

(1) Definition and development of the metadatabase. The definition of the metadata element set (the OmniPaper Application Profile) to be used in the descriptions of the articles was based on the Dublin Core Metadata Terms (DCMT) [3]. This choice stems from its rich semantic interoperability; it is an ISO standard (15836:2003) [8]; it is an ANSI/NISO standard (Z 39.85-2001) [9]; it has been a stable element set since 1996; it has been a Dublin Core Metadata Initiative (DCMI) recommendation since 1999 with its version 1.1; and finally due to the fact that is widely used metadata element set across boundaries of disciplines or application domains [1]. However the DCMT didn't contain all the necessary elements in the description of the news articles. Therefore a new namespace RDF Schema (called "omni") had to be defined in order to add the metadata elements established by the consortium partners [11].

(2) Definition and development of the conceptual layer (subject + lexical thesaurus). The goal of the conceptual layer was to have an ontology-based web of concepts linked to the articles [1]. The used solution was the International Press Telecommunications Council Subject Codes (IPTC SC). The IPTC SC is composed of a hierarchical three-level tree of subject codes, which describes the content of a set

of concepts and has no associative relationship between its concepts. To represent the IPTC SC, several ontology languages were analyzed and studied in order to select the one that best fits its hierarchical representation. However the IPTC SC in the semantically point of view is not that rich and, due to its simplicity, it was only necessary to define its hierarchical concepts. Therefore the RDF Schema language was chosen to complete the representation of the hierarchical tree represented in the IPTC SC [10].

The IPTC SC was then included in a metadatabase. The connection with the subject elements included in the hierarchical tree of the IPTC SC is made through the metadata element "dc:subject". Furthermore, in the OmniPaper application profile definition, the "rdfs:range" of the metadata element "dc:subject" are the IPTC Subject Codes. This means that the metadata element "dc:subject" only allows values originating from the IPTC SC [10].

In addition to the navigation and browsing functionalities through the concepts represented in the IPTC Subject Codes structure, another empowering information-organization tool was included and linked to the metadatabase: WordNet® [1]. The WordNet is "an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets" [12].

An RDF-encoded version of the WordNet 1.6 was downloaded and stored in a local metadatabase. Its connection with the metadata elements of the articles stored in the metadatabase is performed through the metadata element "omnikeyList". This feature enables the user to perform query expansion and thus refine the search term introduced.

The system design of the OmniPaper RDF prototype is shown in figure 1.

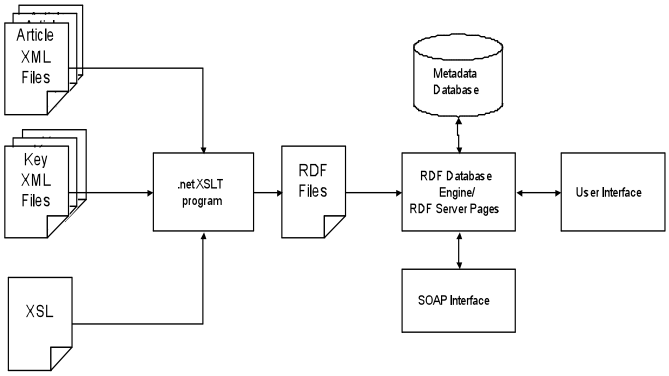


Fig. 1 OmniPaper RDF Prototype System Design

The combination of the subject thesaurus (IPTC SC) with a lexical thesaurus (WordNet), implemented in the OmniPaper RDF prototype to enhance both user queries and navigation is considered rare within RDF applications [1]. Moreover, the fact of these search and browsing functionalities, through the use of a subject thesaurus and a lexical thesaurus being supported by central RDF metadatabase, instead of the full text search, directly in the local archive, usually increase the response time and the results have low levels of precision and recall to the user.

In order to take advantage of the search and navigation functionalities and the whole RDF infrastructure developed in the OmniPaper RDF prototype, it was implemented a system which is an instance of the OmniPaper RDF prototype, in the context of scientific publication. The following sections present the work developed in the instantiation of the OmniPaper RDF prototype in the context of the scientific contents.

3 Instantiation of the OmniPaper RDF Prototype in the Context of Scientific Literature

The smart search mechanisms developed in the OmniPaper RDF prototype, previously presented, demonstrate that the OmniPaper RDF prototype goes far beyond the current full-text search methods. An instance of this prototype in the context of scientific literature aims to improve the daily work of scholars providing sophisticated mechanisms to access and search for new scientific contents.

The implemented system follows the concept of the OmniPaper RDF prototype. However the OmniPaper system was developed in the scope of the news published in newspapers, and its instantiation in the context of scientific publications required some changes at the level of the data and of some processes. In fact, the metadata element set was the principal change in the instantiation process of the OmniPaper RDF prototype, as a consequence of the specific features that characterizes the scientific contents. Another difference stands for the RSS approach followed in the definition of the metadata layer, while in the OmniPaper RDF prototype the metadata description followed the RDF/XML technology. However this can not be considered as a relevant difference because the RSS version 1.0 used, conforms the W3C's RDF specification and is extensible via XML-namespace and/or RDF based modularization [2]. The extensibility mechanisms built in to RSS 1.0 ensure the semantically rich metadata description of any resource which can be identified by a URI. For this reason, the RSS 1.0 was used for metadata encoding and also because it increases interoperability with other RDF/XML applications and enables the rich semantic description of Web resources. However, in the OmniPaper RDF prototype an RDF file holds one description about one article/resource that is stored in one of the local archives, while in the RSS approach is defined a RSS feed that contains a list of items containing titles, links, descriptions and other terms of a set of scientific articles. The differences between the two approaches are in metadata elements used in the description of the Web resources.

Finally, the last difference is related to the navigation functionality available in both systems. In the system developed in the scope of scientific literature it was considered more adequate to use the controlled vocabulary ACM CCS instead of the IPTC Subject Codes used in the OmniPaper RDF prototype. The use of the WordNet remained because it is a lexical thesaurus oriented to organize information, and therefore its use is not restricted to any specific kind of information resource.

4 Implementation of the Metadata Layer

As in the OmniPaper project, the principal aspect of the implemented system was the definition of the metadata layer. The purpose of the metadata layer is to support the search and navigation functionalities and facilitate the access to the contents presented within the feed repository. Another goal concerns the use of the central metadatabase to accomplish the syndication of contents, through the RSS approach. Providing the RSS feed to scholars, they can subscribe the feed and thus be aware of new scientific articles published in the distributed repositories, instead of accessing the different information sources to search for new contents that have been published.

The consumer of a scientific feed requires more information besides the regular metadata elements, usually used in the RSS documents as title, link and (optionally) description. In fact, the user requires more information about the article, in order to be able to cite, or produce a citation for a given article within a serial [5]. Consequently rich metadata was necessary in describing the article along with the article title, link and description. Thus, taking into account that the RSS 1.0 is based on RDF, makes the RSS 1.0 increasingly attractive to be used in the description of scientific publications and therefore ideally suited to the inclusion of supplementary metadata [6]. Several metadata standard vocabularies widely used in the domain of the scientific literature were analyzed in order to define a set of a metadata elements which best describe the features of the scientific contents. The vocabularies were analyzed in terms of the semantics of their elements, their usage in the community, and their interoperability across communities. This analysis resulted in the selection of the DCMT vocabulary in the definition of the metadata structure, since it provides a variety of metadata elements that conforms totality all the necessary descriptions requirements of the information resources used, it enables the metadata interoperability, and it is an ISO and an ANSI/NISO standard.

The selected elements were included in the application profile and the rules for metadata encoding were defined accordingly to the structure defined in the RSS template.

The encode of the metadata elements followed the recommendations made by Kokkelink and Schwänzl in the document “Expressing Qualified Dublin Core in RDF/XML” [7], although, it is still a DCMI candidate recommendation.

5 Implementation of the Conceptual Layer

The design of the system implemented in the context of scientific publications (shown in figure 2) is the same developed in the OmniPaper RDF prototype. The whole RDF model developed in the OmniPaper was instantiated in context of scientific literature. The changes performed were at the data level and in some processes.

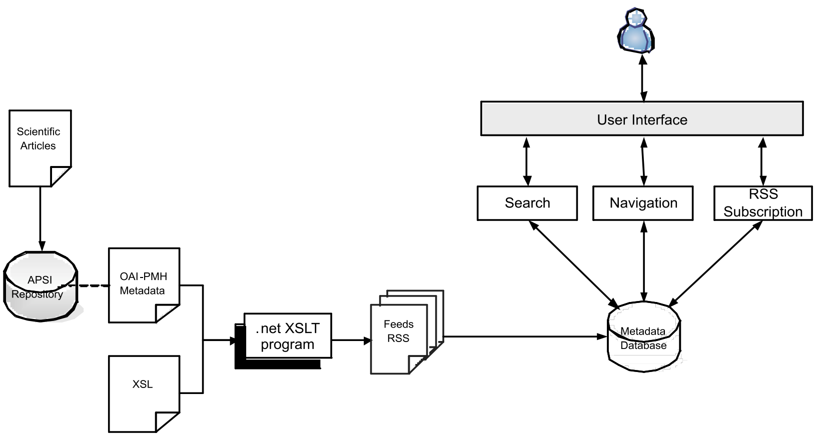


Fig. 2 System Design

The scientific contents used in the developed system are provided by the Associação Portuguesa de Sistemas de Informação (APSI) [13]. APSI has an institutional repository that stores, preserves, disseminates and enables access to the articles published in the Information Systems journal and articles published in the APSI’ conference (CAPSI) proceedings.

APSI provided the metadata information of the articles in the XML OAI-PMH format. Then, in order to get these metadata information encoded accordingly to the RSS structure defined in the template, a stylesheet to perform the transformation was developed. This process resulted in the creation of the RSS feed organized by a list of items. Each item contains the metadata description of one article. RDF triples were extracted from the RSS feed and stored in the metadatabase, through the use of the RDF Gateway, a Microsoft Windows based native RDF database management system combined with a HTTP server. Some RDF Server Pages (RSPs) were defined in order to provide some functionally for the end user.

The metadatabase supports the navigation and browsing functionalities enabling the user to search through the metadata layer and not directly in the information source. In the development of the navigation mechanism it was used the available RDF-S version of the ACM CCS [14]. The scientific articles provided by APSI weren't classified neither using the ACM CCS nor any other controlled vocabulary, thus this work still had to be done manually. The connection between the ACM CCS subject tree to the metadatabase is performed through the "dc:subject" metadata element, as illustrated in figure 3. This fact allowed the user to subscribe the articles within a specific subject accordingly with the specific interest areas of each subscriber.

Besides this functionality, the use of the WordNet was also instantiated from the OmniPaper RDF prototype. The use of this lexical thesaurus improves the search procedure, since it allows the relationship of the input concept with others, enabling the user to perform query expansion and thus redefine his search. The connection to the articles' descriptions stored in the metadatabase was made through the "description" metadata element, as illustrated in figure 3.

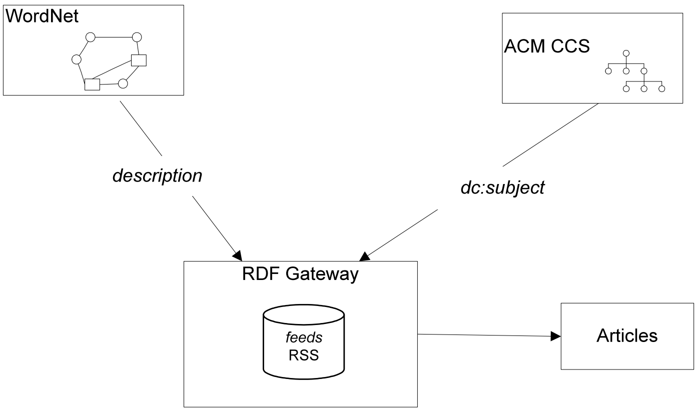
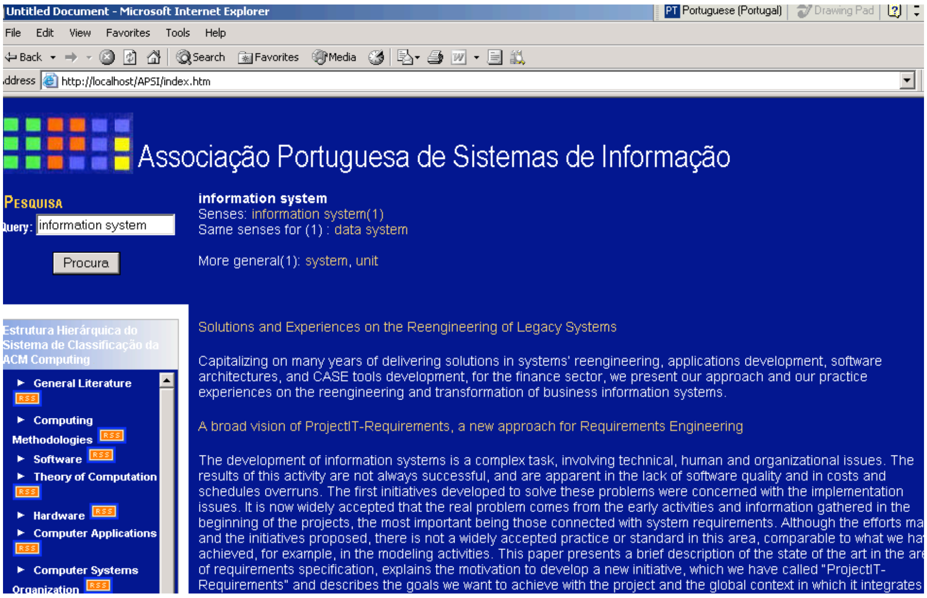


Fig. 3 Metadata semantic layer of the system

Furthermore, the metadatabase also enabled the syndication of scientific contents that are included in the RSS feed. In fact, the user doesn't need to check for new articles published in the APSI repository, because since the user subscribed the RSS feed he can be aware of new issues that have been published through an RSS Reader application with which the user is familiar with.

Figure 4 shows the interface of the implemented system.


Untitled Document - Microsoft Internet Explorer
PT Portuguese (Portugal)
Drawing Pad
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
ddress
http://localhost/APSI/index.htm
Associação Portuguesa de Sistemas de Informação
Pesquisa
Procura
Estrutura Hierárquica do Sistema de Classificação da ACM Computing
General Literature RSS
Computing Methodologies RSS
Software RSS
Theory of Computation RSS
Hardware RSS
Computer Applications RSS
Computer Systems Organization RSS

uery: information system

information system
Senses: information system(1)
Same senses for (1) : data system

More general(1): system, unit

Solutions and Experiences on the Reengineering of Legacy Systems

Capitalizing on many years of delivering solutions in systems' reengineering, applications development, software architectures, and CASE tools development, for the finance sector, we present our approach and our practice experiences on the reengineering and transformation of business information systems.

A broad vision of ProjectIT-Requirements, a new approach for Requirements Engineering

The development of information systems is a complex task, involving technical, human and organizational issues. The results of this activity are not always successful, and are apparent in the lack of software quality and in costs and schedules overruns. The first initiatives developed to solve these problems were concerned with the implementation issues. It is now widely accepted that the real problem comes from the early activities and information gathered in the beginning of the projects, the most important being those connected with system requirements. Although the efforts ma
and the initiatives proposed, there is not a widely accepted practice or standard in this area, comparable to what we ha
achieved, for example, in the modeling activities. This paper presents a brief description of the state of the art in the are
of requirements specification, explains the motivation to develop a new initiative, which we have called "ProjectIT-Requirements" and describes the goals we want to achieve with the project and the global context in which it integrates

Fig. 4 Screenshot of the developed prototype

6 Conclusions and Future Work

The development of information and communication technologies, and the increase use of the Internet associated to the user needs have been contributing to a deep restructure of the traditionally means used in the publication of information. The RDF prototype implemented in the OmniPaper project has already proven to be an efficient system to search for news contents published in newspapers. Moreover the RDF application combines completely with subject ontology and a lexical thesaurus to enhance both user search and navigation. In this context, the instantiation of the OmniPaper RDF prototype in the scope of the scientific publication would improve the current mechanisms used to access and to distribute the scientific research developments. This article described the steps followed in the instantiation of the OmniPaper RDF prototype in the context of scientific publications. The main difference between the two systems was the use of the RSS technology in the metadata description, enabling the syndication of the scientific feeds.

As future work it is necessary to perform the evaluation of the system implemented, to determine the relevance of the results returned and its usability. It would also be interesting to harvest several repositories to the metadatabase implemented in this system, in order to provide a more complete service with more information.

References

1. Baptista, A. A.: Searching and browsing using RDF-Encoded Metadata: the case of OmniPaper. Canadian Journal of Communication. 29 (3), 317--328 (2004). Available from: https://repositorium.sdum.uminho.pt/handle/1822/5080.

2. RDF Site Summary (RSS) 1.0, http://web.resource.org/rss/1.0/spec#.

3. Dublin Core Metadata Element Set, Version 1.1: Reference Description, http://www.dublincore.org/documents/dces/.

4. Dublin Core Metadata Initiative Home Page, http://www.dublincore.org/.

5. Hammond, T.: Why Choose RSS 1.0? XML.com (2003). Available from: http://www.xml.com/pub/a/2003/07/23/rssone.html.

6. Hammond, T., Hannay, T. e Lund, B.: The Role of RSS in Science Publishing Syndication and Annotation on the Web. D-Lib Magazine, 10 (12) (2004). Available from: http://www.dlib.org/dlib/december04/hammond/12hammond.html.

7. Kokklink, S. e Schwänzl, R.: Expressing Qualified Dublin Core in RDF/XML [on-line]. Dublin Core Metadata Initiative (2002). Available from: http://www.dublincore.org/documents/2002/04/14/dcq-rdf-xml/.

8. National Information Standards Organization. The Dublin Core Metadata Element Set: An American National Standard/ developed by National Information Standards Organization (2001), http://www.niso.org/standards/resources/Z39-85.pdf

9. National Information Standards Organization. NISO Press Release - The Dublin Core Metadata Element Set Approved (2001), http://www.niso.org/news/releases/PRDubCr.html

10. Pereira, T. and Baptista, A. A.: Incorporating a Semantically Enriched Navigation Layer Onto an RDF Metadatabase. Engelen, J., Costa Sely., M. S., Moreira, Ana Cristina S., ed. In: Building digital bridges: linking cultures, commerce and science: Proceedings of the ICCC International Conference on Electronic Publishing, ELPUB, July 2004 Brasilía, Brasil (2004). Available from: https://repositorium.sdum.uminho.pt/handle/1822/604.

11. Pereira, T., Yaginuma, T. and Baptista, A. A.: Perfil de Aplicação e Esquema RDF dos Elementos de Metadados do Projecto OmniPaper. In: CMLE'2003 – 3th Congresso Luso-Moçambicano de Engenharia. August 20, 2003. Maputo, Moçambique (2003).
Available from: http://repositorium.sdum.uminho.pt/handle/1822/281

12. Princeton University Cognitive Science Laboratory. WordNet –A lexical database for the English language, http://www.cogsci.princeton.edu/~wn

13. Associação Portuguesa de Sistemas de Informação, http://www.apsi.pt/.

14.ACM Computing Classification System [1998 Version], http://dspace-dev.dsi.uminho.pt:8080/pt/addon_acmccs98.jsp

15. Paepen, B.: Blueprint: a universal standard model for efficient information retrieval. Technical Report of the OmniPaper Project, 28 February of 2005. Available from: http://www.omnipaper.org/