Information on the Omnipaper project

Document history				[Document URN: urn:iptc:workdoc:dir:0401:1]
Revision	Issue Date	Pages	Author (revised by)	Remark
Unrevised	2004-01-09	22	Michael Steidl, MD IPTC	
Rev a.				

A project on assigning metadata element to existing newspaper articles – named "Omnipaper Project" – is conducted at the University of Minho in Portugal. One of the project leads, Ms Teresa Susana Mendes Pereira, was in contact with the IPTC Managing Director since 2002 to receive the IPTC documentation on NITF and NewsML and joined the respective IPTC discussion groups in 2003 to learn more about. She focussed her assessment on NITF and NewsML and sent requests for additional and background information either to the MD or the Yahoo newsgroup.
Now papers of the Omnipaper Project were sent – from another source – to the IPTC office showing the status of this project as of summer 2003 (that was inferred from document properties of the document files).

This package shows two documents:

- Design of metadata elements for digital news articles in the Omnipaper Project
- The Ominipaper metadata RDF/XML prototype implementation

Primarily the first document provides an evaluation of the IPTC standards NITF and NewsML, so this is a good reflection of the perception of our work outside the IPTC community.

(Remark: sorry, but some graphics show a garbled descriptive text after converting the Word document to PDF, apparently embedded Visio graphics is the reason for this error.)

=== END of document ===

DESIGN OF METADATA ELEMENTS FOR DIGITAL NEWS ARTICLES IN THE OMNIPAPER PROJECT

TOMOKO YAGINUMA; TERESA SUSANA MENDES PEREIRA;
ANA ALICE BAPTISTA

Department of Information Systems, School of Engineering, University of Minho
Campus de Azurém, 4800-058, Guimarães, Portugal
e-mail: tomoko@dsi.uminho.pt; tpereira@dsi.uminho.pt; analice@dsi.uminho.pt

This paper examines and proposes a set of metadata elements for describing digital news articles for the benefit of distributed and heterogeneous news resource discovery. Existing digital news description standards such as NITF and NewsML are analysed and compared with Dublin Core Metadata Element Set (DCMES), which results in that the use of Dublin Core is encouraged for interoperability of the resources. The suggested metadata elements are carefully selected and defined considering the characteristics of news articles. Some elements are detailed with refinement qualifiers and recommended encoding scheme. This set of metadata has been developed as a part of the tasks in the IST (Information Society Technologies)-funded European project OmniPaper (Smart Access to European Newspapers, IST-2001-32174).

Keywords: metadata, resource description, Dublin Core, digital news, interoperability, OmniPaper

1. INTRODUCTION

It has become inevitable to use metadata for resource description in order to enhance information retrieval from distributed and heterogeneous resources. To search in semantically and syntactically common set of metadata elements makes resource discovery easier and more efficient.

A number of studies on metadata standards have been carried out in different application domains such as annotation of images [1], internet resources retrieval [2] and agricultural resource cataloguing [3]. The IST-funded OmniPaper project [4] investigates ways for smart access to news articles originating from a large number of European digital newspapers. Thus, there is a need for definition of the common set of metadata elements that provides a uniform entrance to the distributed news articles.

This paper analyses possible metadata description standards that might be relevant to describe news articles and then proposes the metadata element set specifically defined for OmniPaper. First of all, aims and steps taken for development of the elements are briefly explained.

Secondly, existing standards for digital news description are analysed and compared with more general metadata description format provided by Dublin Core Metadata Initiative (DCMI) [5]. Then, a set of core 25 elements defined for OmniPaper will be introduced with explanation on some remarkable elements, followed by suggested future work. Finally, conclusion will be drawn.

2. NEED FOR THE COMMON METADATA ELEMENT SET: AIMS AND STEPS TAKEN FOR IMPLEMENTATION OF METADATA SET

In most European countries, initiatives do exist or are at least initiated for newspaper article exchange on a larger scale. These initiatives all share some limitations [6]:

(1) They all use a very centralized approach, that is, newspaper-articles are sent in a more or less standard format for check-in in a central database system that resides at a service provider's site.

(2) Most of these initiatives do not cross language or country boundaries.

OmniPaper project intends to overcome both limitations. The first step taken is to create a common standardised interface to distributed local news archives situated in different European countries, which are all within different operating environments, database formats and indexing mechanisms. Most of the resources are already digitised in XML format but structured in different schemas. The OmniPaper system transforms the resource documents into the common text format structured with News Industry Text Format (NITF) [7], using XSLT. SOAP will be used to correspond between local archives and the OmniPaper system. This means that each local news provider does not need to make any change on its text format. Simplified diagram shown in figure 1 describes the transformation architecture.

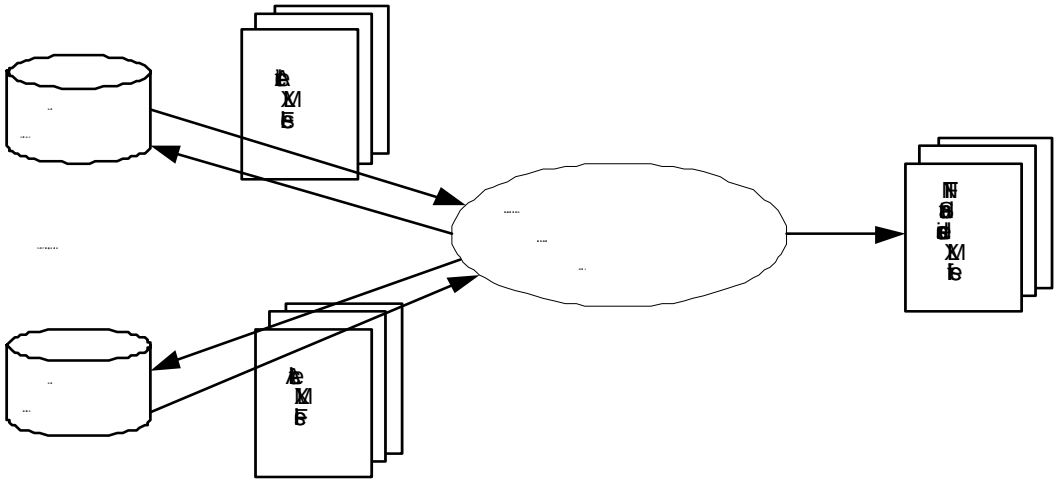

FIGURE 1 STANDARDISED TEXT FORMAT TRANSFORMATION

Besides the text transformation, metadata contained in the original documents will be extracted and described in metadata description technologies. OmniPaper uses Resource Description Framework (RDF) [8] and XML-Topic Maps (XTM) [9] to describe these extracted metadata and implement a prototype for each. With these technologies, multilingual aspects of resources will be handled in a well-organised manner, too.

In order to set up a uniform interface to the different resources, these metadata documents independent either from original news documents or transformed NITF documents must have the common set of metadata elements. This set of metadata should be able to represent each news article in a simple and well-structured way so that users can search on metadata to find their information requests. Obviously, the OmniPaper system will contain several search options including full text search on the content of the news articles and cross search on the different metadata properties, etc.

From all of the above, definition of core metadata element set is a critical task in the early stage of the project. To execute this task, several steps are planned and taken: (1) study and survey on existing metadata standards and the metadata structures of our partners (news providers); (2) creation of initial element set; (3) revision with the project partners; (4) refinement of the initial set (5) definition of encoding schemes for appropriate elements; (6) implementation of final set of metadata.

In the next section, we will present our principal study on news format standards largely used in the news industry in order to develop the OmniPaper metadata element set.

3. ANALYSIS OF STANDARDS FOR METADATA ELEMENTS DESIGN

Currently, there are three global news format standards widely used in the news industry, that are NITF, NewsML [10] and XMLNews [11]. The first two are implemented and maintained by the International Press Telecommunications Council (IPTC) [12], that is an organisation developing and publishing industry standards for the interchange of news data. The last is developed by the XMLNews.org[1].

NITF is a text format standard which defines the content and structure of news articles [12]. It was first created based on the Standard Generalized Markup Language (SGML) and then after 1998 up-dated to be compliant with the Extensible Markup Language (XML). NITF supports the identification and description of a number of news characteristics with rich in-line markups. It identifies, for instance, structural pieces of news article such as headings, bylines, paragraphs, tables, columns and footnotes. Metadata tags are also applied for describing the news content.

NewsML is a one level higher standard for news representation and management [12]. It is also based on XML. The main function of NewsML is the management of news item throughout its lifelong, including production, interchange and consumer use. It can contain many different types of objects including multimedia items such as images, videos, etc. The text written in NITF can be accommodated within NewsML in self-contained manner. Likewise, other formats like SportML and JPEG can be wrapped by NewsML. While NITF is a standardised format for description and structuring of news text, NewsML is represented as an information wrapper and management tool.

XMLNews is a pair of specifications, XMLNews-Story which defines the content of textual news stories, and XMLNews-Meta which defines a set of metadata information about news objects. Both specifications are developed based on XML in 1999. XMLNews-Story is a fully-compatible subset of the NITF. XMLNews-Meta is a news industry metadata format conforming to RDF. It is used as a single, common format to provide metadata about any kind of news object, whether textual or non-textual, separated from the news objects themselves.

Unlike XMLNews, NITF and NewsML do not have metadata elements separated from news content description. In other words, metadata is included in the same file of the news content, or located in the several places within a certain news object's structure. However, as the efficient use of metadata is one of the most important features of these standards, we studied the use and features of metadata, and found out that there are several remarkable points to characterise these metadata.

First of all, the diversity of the elements is striking. Both NITF and NewsML have a number of elements that can be used to describe details of news article content, format, copyright, management, etc. NewsML categorises them such as AdministrativeMetadata, RightsMetadata, DescriptiveMetadata and PublishingMetadata. Most of the elements are designed to be as close as possible to the items they describe, while much of the metadata is optional. It is also worth noting that NewsML provides a uniform metadata for all media types.

Another notable point is that IPTC handles controlled vocabulary for both standards outside the standards' DTD. The famous one is IPTC Subject Reference System, which is a hierarchical three-level tree of subject codes. With the use of this system for a certain metadata, the subject area covered by the content of a news article can be indicated in a uniform way. It is extremely beneficial and efficient to use this kind of standard vocabulary for metadata encoding as it increases interoperability of news articles.

Lastly but not least, there is a strong feature for news management, especially in NewsML's metadata. Some of the metadata is well designed to be able to add or replace the

[1] Maintained by Megginson Technologies Ltd.

values of metadata easily over the lifetime of an article. This flexibility is highly suitable for dynamic aspects of news. Moreover, some metadata allow evaluation of metadata when necessary and appropriate. News items are not dealt with only one person or one company. Normally they are passed through several processes and in each stage new metadata could be added. After news items are published, there may be a third party which revises or evaluates it. Metadata for evaluation is useful to update the status of news items, record the history of them, and clarify the entity which assigned the values.

In terms of XMLNews, it appears to be suitable for the OmniPaper purpose in a way that it can provide metadata separately from news content. Compatibility to RDF seems to be beneficial for implementation of the RDF prototype. More than 40 metadata elements defined is also enough to describe the news object.

Considering all the above, it sounds good enough to use either NITF, NewsML or XMLNews as they are to describe the OmniPaper news articles, including the metadata elements. In fact, NITF standard will be used as a uniform text format for the OmniPaper-participating news articles in favour of further process in the OmniPaper system. However, there are some disadvantages in using these existing standards for dealing with metadata in the following context.

The documents written in NITF and NewsML standards contain not only metadata information but also the entire news content in the same file, which is not necessary for metadata document in RDF or XTM. Search on metadata will be executed much faster and easier if the system will be able to access only metadata in the separate files. If the document contains the content of news or other elements not necessarily used for search, it will only consume a disk space. In addition, the structure of NITF and NewsML formats are rather complicated. It is not easy to locate a certain metadata, both for human use and machine use. Consequently, to use the metadata elements from NITF or NewsML as combined with the article content is not in the line of the OmniPaper purpose.

XMLNews was further examined but it turned out to be less popular than NITF and NewsML. This tendency might stem from the fact that XMLNews-Story is originated from NITF, and NITF has stronger support from users as well as NewsML.

As a result, OmniPaper in principle does not use these standards for metadata description and handling although they were largely studied as a reference. Instead, our choice for metadata description went to Dublin Core Metadata Element Set (DCMES) [13] developed by DCMI.

DCMI is an organization which main focus is to develop semantic metadata and specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems [5]. Currently, fifteen core elements are defined as DCMES. DCMES is relatively simple metadata set, easy to understand and extend to richer semantic description standards. Moreover, it has been widely used across the boundaries of disciplines or application domains.

One of the benefits to use DCMES stems from this semantic interoperability. It is a simple and general standard but it makes the barrier of domains much lower. In other words, news articles described with DCMES have higher possibility to be discovered by external applications in different area, not only in news industry. This achieves Tim Berners Lee's Semantic Web concept [14].

For the reasons above, DCMES is taken as OmniPaper's principal metadata standard vocabulary. Furthermore, we agreed on the priority that we should follow when no appropriate element is found in DCMES. Based on the idea that metadata should be described using as many standards commonly used as possible, the next priority should go to NITF, NewsML or XMLNews. The last option is OmniPaper-original metadata, that is the creation

of elements. This must be avoided unless there is strong necessity. Figure 2 summarises metadata vocabularies or DTD already established for the OmniPaper.

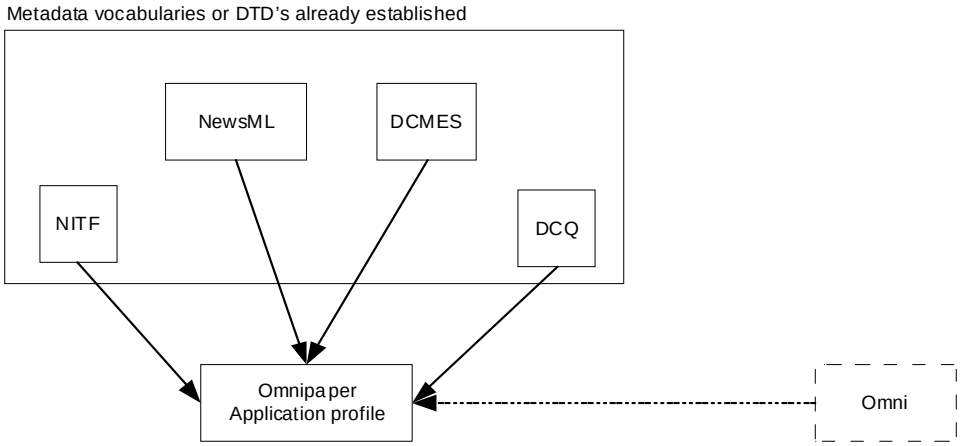

FIGURE 2 METADATA VOCABULARIES ESTABLISHED FOR OMNIPAPER

Along with the study of existing text format standards, we have examined the metadata elements that our news provider partners currently use for their systems. It is highly important to know what metadata elements are in use in the real world. Close examination showed several characteristics of the elements for describing news articles.

(1) There are some common and essential metadata that all the companies use. They refer to the title of an article, author, issue date and the id for the article.

(2) There are some kind of classification elements which sort the articles into categories, or reference elements which link the articles related to each other or in the same series. Each company has its own way of classifying and linking articles, using an existing or original thesaurus or other criteria.

(3) One company has several elements which describe the photos used for an article in details. Another company has the element which only describes the caption of the photo. These elements brought out the issue of how to deal with other media included in the news articles.

(4) Some elements are highly based on the paper version of news articles. For instance, they describe from which edition(s) the article came from, or in which section of the newspaper the article belongs to.

These factors were greatly useful for the selection of metadata elements. According to the analysis described in this section, our initial metadata element set was created. The following section introduces the final version of OmniPaper metadata element set.

4. OMNIPAPER METADATA ELEMENT SET: THE MODEL FOR DIGITAL NEWS RESOURCE DESCRIPTION

We have defined 25 elements for the OmniPaper core metadata set. The tables below show all the elements sorted into 6 categories: Article Identification; Article Ownership; Article Location (Storage); Article Relevance/Audience; Article Classification; Link Information. Each element is described with name, definition of the element, the source of the element and recommended encoding scheme(s). The encoding schemes are the qualifiers used for expressing the element value. These schemes include controlled vocabulary and formal notation. For the metadata elements taken from DCMES and Dublin Core Qualifiers (DCQ) [15], we followed the encoding schemes recommended by DCMI.

As can be seen in the Source column, 15 elements are defined based on Dublin Core, either DCMES or DCQ, and 7 elements are based on NITF or NewsML, and the rest 3

elements are OmniPaper-original. Majority of elements are taken from Dublin Core, which is ideal for maximising interoperability. All the elements are repeatable.

TABLE 1 - ARTICLE IDENTIFICATION

Source	Element	Definition	Encoding Scheme(s)
DCMES	Identifier	An unambiguous reference to a resource (an article).	URI
	UniqueId	An OmniPaper specific ID for a resource (an article).	
DCMES	Creator	Author(s) of an article.	
DCQ	Issued	Date of publication of an article.	W3C-DTF
DCMES	Title	Title of an article.	
NITF	Subtitle	Subtitle of an article (if any).	
DCMES	Publisher	The entity which an article belongs to.	
DCMES	Language	The language in which an article is written.	ISO 1766 & 639
	KindOfArticle	Nature or genre of an article.	
NITF	Section	Named section of a publication where an article appears.	
NITF	Edition	The name(s) of edition(s) in which an article is distributed.	

TABLE 2 - ARTICLE OWNERSHIP

Source	Element	Description	Encoding Scheme(s)
NITF	Copyright	Container for copyright information.	
	Supplier	The local archive that owns the article.	

TABLE 3 - ARTICLE LOCATION (STORAGE)

Source	Element	Description	Encoding Scheme(s)
DCQ	Medium	The physical or digital manifestation of an article.	IMT
DCMES	Source	A reference to an article from which the present article is derived.	URI

TABLE 4 - ARTICLE RELEVANCE/AUDIENCE

Source	Element	Description	Encoding Scheme(s)

NewsML	OfInterestTo	The target audience for an article, based on demographic, geographic or other groups.	
DCQ	Valid	Date (often a range) of validity of an article.	DCMI Period, W3C-DTF
DCQ	Spatial	Geographical location that an article treats or is related to.	DCMI Point, ISO 3166, DCMI Box, TGN

TABLE 5 - ARTICLE CLASSIFICATION

Source	Element	Description	Encoding Scheme(s)
DCQ	Abstract	A summary of the content of an article.	
NITF	Key_list	A container for list of most relevant keywords extracted from an article document.	
DCMES	Subject	Topic of the content of an article, specified according to the common thesaurus.	IPTC Subject Code System, etc.

TABLE 6 - LINK INFORMATION

Source	Element	Description	Encoding Scheme(s)
DCQ	HasPart	The described article includes the referenced resource such as photo, table, diagram, etc	URI
DCQ	IsVersionOf	The described article is a version of the referenced version	URI
NITF	Series	The name of series that an article belongs to	
DCQ	References	The described article references, cites or points to the referenced article	URI

Some additional explanation is given to the elements below.

UniqueID:

UniqueID is an OmniPaper-original element. The purpose of this element is to identify a certain article with OmniPaper specific string-based ID for the process within the OmniPaper system. The use of NITF element called id-string was considered as it defines “character string that provides a unique, persistent identification for a document”. However, the element is internal use only and does not need to be public so that the original element was selected.

KindOfArticle:

This element is also an original of OmniPaper. In NITF, there is one element which defines type of article such as analysis, feature, obituary and so on. To be exact, this element

is the attribute of “tobject.property” element, and not very clear to use in this way. Therefore, KindOfArticle remains as original.

Edition:

The element Edition has to identify regional versions of an article. In general, an article has different versions for different regional editions of the same newspaper. The same version of an article can appear in the different regional editions of the same newspaper, too. Therefore, Edition may have one or more values.

Supplier:

This element has to describe the organisation which supplies news articles to the OmniPaper system. It can be called “Local archive” as conventionally used in the OmniPaper architecture and consortium. It has to be clearly distinguished from the author of the article, news companies or distributors which provide news to the “local archive”. NITF Distributor was one of the candidates but we took the original “Supplier” to avoid the confusion of the specific meaning.

Medium:

In the early stage of the development, we had the elements Format (from DCMES) and MediaType (from NewsML) for describing the format of an article and other media included in the article. MediaType elements are typically expressed as text, graphic, photo, video and so on. We realised that this metadata would be better described by the DCQ element Medium. This way, the metadata was refined with the qualifier which makes the meaning of the element clearer and more specific.

Source:

The Source element is used to reference to a source article when any other documents are created based on it. For example, if there is an opinion article which talks about the other article, the metadata document for the opinion article will reference the target article with the Source element.

Valid:

This element can be used to describe the lifetime of an article typically with “start date” tag and “end date” tag. The expired-dated article may be removed from the database.

Spatial:

Previously, there were several more elements which refer to the semantic content of the document, namely, Doc_scope (NITF), Location (NITF), Company and Organisation. Doc-scope indicates an area where the article may be of interest. Location defines a relationship between an article and a geographical location about which the article treats. Company and Organisation define the relationship between an article and a company/organisation about which the article treats. The first two are very similar so that we had to have only one element which describes the geographical space. Then, the Spatial element, the qualifier of Coverage (DCMES) was chosen instead of NITF based Doc_space and Location. The latter two elements are rather too specific and we decided to discard them.

Subject:

The Subject element is used to classify articles based on the common thesaurus in the OmniPaper system. One of the candidates is the IPTC Subject Code System, which classifies the subjects in three levels. It is possible to assign one or more subjects to the same article.

HasPart:

The element HasPart is useful when an article contains not only text but also photo, table, diagram, etc. These resources within the article are described as the “part of” the article and the HasPart element points to each resource using URI.

IsVersionOf:

Normally, an article has only one version. However, if new information for the article was obtained or mistake was found even after printing process had begun, the article is often

modified. This new version will contain the element isVersionOf, which points to the original article.

Series:

A series of articles are often published over a longer period covering the same topic. The element Series can tie these articles together and make the retrieval easier and more precise.

References:

The References element relates an article to the other article. Possible cases are when an article refers to other articles for more detailed information or when an article is continued to several pages.

These core elements are now described in XML/RDF, and the application profile is also created [16].

5. FUTURE WORK

There are several elements which may need some refinement or a decision of which encoding scheme should be used. These elements are: KindOfArticle; Section; Copyright; OfInterestTo; Spatial.

The element KindOfArticle aims at classifying the form of an article as genre. It is strongly related to Type element in DCMES, which definition is “the nature or genre of the content of the resources”. Therefore, KindOfArticle could be defined as the sub-properties of DC Type.

There is a controlled vocabulary for the values of Type property, called DCMI Type Vocabulary [17], which includes Dataset, Event, Sound, Interactive Resource, etc. They are still too general terms and not exactly suitable for describing a specific genre like the form of the news article. Consequently, we will need to define our own set of vocabulary, which might be used internally.

For refinement of Copyright element, additional attributes such as the date of the copyright ownership established and the name of copyright holder will have to be included. We had more specific elements such as RightsEndDate for specifying the duration of copyright validity or RightsGeography for spatial coverage of the rights in the early stage of discussion. However, they are not included to this report as most of them are not applicable to the nature of news articles. The Copyright element can be “subPropertyOf” DC Rights element, which is a broader category for copyright, intellectual property rights and so on.

OfInterestTo element is one of the most challenging but interesting elements in the OmniPaper metadata set. In terms of target audience, there is a working group in DCMI within which the discussion of the definition of DC Audience element and its audience levels are ongoing [18]. Since the qualifiers proposed are mainly related to educational domains, it may not be relevant to describe the audience level of news subscribers. We will need to study user profile when we define the controlled vocabulary.

Our next step will be implementation of transformation module, which works as a mapping of existing metadata elements in local archives to the OmniPaper metadata elements. As there is no strict obligation for the definition of elements for the news providing companies yet, this mapping process first has to be done with studying each local archive’s metadata elements’ characteristics, encoding schemes and structure. Then, each element will be mapped onto the relevant element in the OmniPaper system.

In order to examine the efficiency of metadata use for search, prototype testing will be necessary. For the OmniPaper project, two metadata description technologies (RDF and XTM) are used. These two prototypes will be compared with the one which uses full text search on the content of the news articles, too.

6. CONCLUSION

In this paper, we analysed three existing news text format standards, which are NITF, NewsML and XMLNews, in order to examine if they can be used to describe metadata of news articles for the benefit of OmniPaper project. There are a number of useful features in these formats. However, they are too complicated in their structures and use. To understand and use them efficiently not only for describing news resources but also for interactive search might be slightly difficult and time-consuming. Besides, our aim is to describe only metadata and there is no need to include original text in the file as NITF and NewsML do. It is also important for us to implement a metadata set fundamental and useful for resulting in better information discovery and interoperability. As a result, more general metadata description scheme DCMES was selected as our basis for implementation of the OmniPaper metadata set.

Choice of DCMES for the base of the OmniPaper metadata set must be ideal as DCMES has been already widely used across the different domains. This leads to greater interoperability in terms of search on Internet. It was also advantageous that DCMES is rather simple and easy to handle.

Having defined 25 core elements for OmniPaper, these elements are covering many aspects of news articles that might result in semantically rich metadata set probably applicable to similar resource applications. Some future works are suggested mainly regarding further refinement on some elements and definition of encoding schemes.

The use of metadata in information discovery will be tested in the OmniPaper prototype testing. It is planned that example queries will be searched in the pool of the documents containing metadata elements. The results will be analysed and further refinement on metadata elements will be made, too.

REFERENCES

1. Campbell, D. Grant, *The Use of the Dublin Core in Web Annotation Programs*, Proc. Int. Conf. On Dublin Core and Metadata for e-Communities 2002, pp.105-110, Florence, Italy, 2002.

2. Neuroth, Heike; Koch, Traugott, *Metadata Mapping and Application Profiles. Approaches to providing the Cross-serching of Heterogeneous Resources in the EU Project Renardus*, Proc. Int'l. Conf. On Dublin Core and Metadata Applications 2001, pp.122-129, Tokyo, Japan, 2001.

3. Onyancha, Irene; Keizer, Johannes; Katz, Stephen, *A Dublin Core Application Profile in the Agricultural Domain*, Proc. Int'l. Conf. On Dublin Core and Metadata Applications 2001, pp.185-192, Tokyo, Japan, 2001.

4. OmniPaper – Smart Access to European Newspapers, EU project IST 2001-32174, http://www.omnipaper.org/, 2002.

5. Dublin Core Metadata Initiative, http://dublincore.org/

6. Schranz, Markus; Tscheligi, Manfred; Paepen, Bert, *OMNIPAPER: modern approaches for an intelligent European news archive*, Paper submitted to ECDL, 2002.

7. News Industry Text Format, http://www.nitf.org/

8. Lassila, Ora; Swick, Ralph R., *Resource Description Framework (RDF) Model and Syntax Specification*, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/, 1999

9. Pepper, Steve; Moore, Graham, *XML Topic Maps (XTM) 1.0*, http://www.topicmaps.org/xtm/1.0/, 2001.

10. News ML, http://www.newsml.org/

11. XMLNews, http://www.xmlnews.org/

12. International Press Telecommunications Council, http://www.iptc.org/

13. Dublin Core Metadata Elements Set, Version 1.1: Reference Description, http://dublincore.org/documents/dces/

14. Berners-Lee, Tim; Hendler, James; Lassila, Ora, *The Semantic Web* http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2, 2001.

15. Dublin Core Qualifiers, http://dublincore.org/documents/dcmes-qualifiers/

16. Pereira, Teresa; Baptista, Ana Alice. *Omnipaper – Arquitectura de metadados e sua implementação no RDF Gateway*, III Congresso Luso-Moçambicano de Engenharia, Maputo, Moçambique, 2003.

17. DCMI Type Vocabulary, http://dublincore.org/documents/dcmi-type-vocabulary/

18. DCMI Education Working Group, http://dublincore.org/groups/education/

THE OMNIPAPER METADATA RDF/XML PROTOTYPE IMPLEMENTATION

TERESA SUSANA MENDES PEREIRA, ANA ALICE BAPTISTA
Department of Information System, School of Engineering, University of Minho, Portugal
Campus de Azurém, 4800-058, Guimarães, Portugal
{analice, tpereira} @dsi.uminho.pt

Abstract

Omnipaper (Smart Access to European Newspapers, IST-2001-32174) is a project from the European Commission IST program (Information Society Technologies) that investigates and proposes ways for access to different types of distributed information sources. This article intends to introduce the technology Resource Description Framework - RDF, developed by W3C for the Web based on metadata, and its practical use in the Omnipaper project, which the authors are involved. We intend to achieve the implementation of a 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. Omnipaper is not a project about digitalization of news, but about bringing digitized news originating from various sources (and in various formats) together. In this article will be described the procedure implemented in the description of our newspaper articles using the RDF technology, followed by a elaborated description on the manipulation process and treatment of the information structured in RDF, through the RDF Gateway.

Keywords: Metadata, RDF**,** DC Web, Metadata.

INTRODUCTION

During the last decades, the amount of digital information has grown exponentially. The same holds for the number of computers and Internet connections. 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 for 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 [, 1999 #26].

One of the important challenges 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 OmniPaper [2] project is investigating ways for drastically enhancing access to many different types of distributed information resources. The key objective of the OmniPaper project is the creation of a multilingual navigation and linking layer on top of distributed information resources in a self-learning environment thus providing a sophisticated approach

to manage multinational news archives with strong semantic coupling, delivering to the user more than the sum of the individual service features.
One of the principal aspects of the project is the whole metadata layer of the system. Actually, there are two metadata layers: (1) the first layer (Local Knowledge Layer) that is added to the local archives where the main purpose 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; and (2) a second layer (Overall Knowledge Layer), an higher abstraction level that will provide a common access user interface by integrating and relating the metadata information coming from the local knowledge layer. Furthermore, this interface is meant to be personalized and multilingual.
This paper explains the research work conducted in the RDF approach to the Omnipaper Local Knowledge Layer and deepens each step shown above.

OVERVIEW TO OMNIPAPER ARQUITECTURE

The Omnipaper's project begun in January 2002 and is structured into seven workpackages, where each one handle a well-defined part of the work, at the same time being highly linked to each other.

As depicted in Figure 1, the architecture used in the project makes a distinction between the Local Knowledge and Overall Knowledge Layer.

The three workpackages included in the Figure 1 (WP2, WP3 e WP5) will each pursue one of the three technological objectives.

On the level of the local layer, ways for retrieving information from distributed sources were analyzed. When combining different archives into one large information pool, access to them must be possible in a uniform way (WP 3). The Overall Knowledge layer will bring to the local layers together in a well-structured manner. It will make cross-archive navigation and linking in a multilingual environment. On top of the overall layer, the user interface will provide a user-friendly and interactive presentation of the knowledge layer (WP 5).

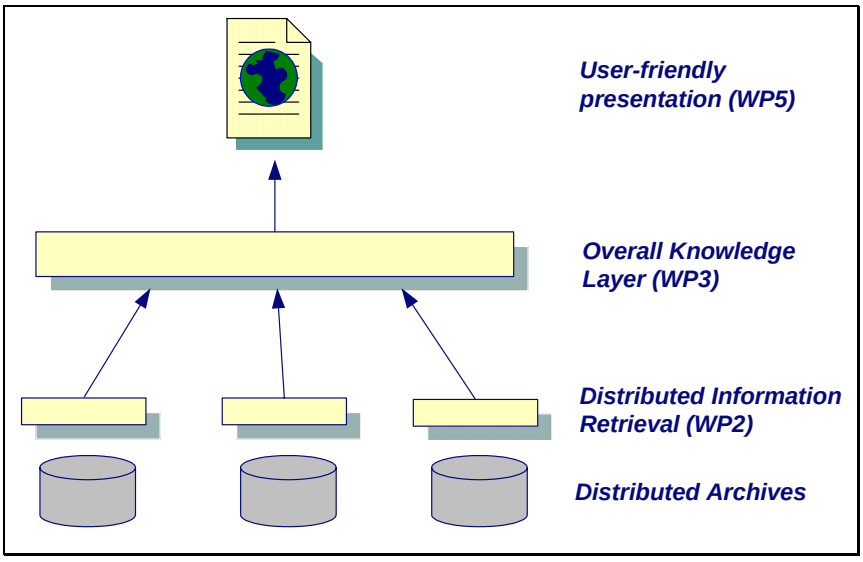

FIGURE 1: SYSTEM ARQUITECTURE

LOCAL KNOWLEGE LAYER

The implementation of the Local Knowledge Layer is developed in the WP2 and results will constitute the input for research on the Overall Knowledge Layer (WP3). On both levels, the RDF and Topic Maps approach will be analysed and compared.
The Local Knowledge Layer is added to the local archives where the main purpose 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.
In the Local Knowledge Layer we are developing simultaneously two different metadata approaches to both layers: the Resource Description Framework (RDF) [3] approach, and the Topic Maps (TM) [4] approach. These prototypes will be cross-tested on a technical level and will allow conclusions on both levels [5].

OVERALL KNOWLEGE LAYER

The overall knowledge layer combines the features of integrating distributed information with the capability of creating semantic coupling of the corresponding content.

The results from the local layer queries will constitute the input for research on the overall layer. On both levels, new techniques are analyzed and compared.

Meta data techniques are evaluated to act as a knowledge management platform in order to improve content research and multilingual information retrieval. The multilingual aspects are supported by extracting existing keywords and metadata from the heterogeneous archive information and associate it with existing domain specific thesauri for the relevant language.

The overall knowledge layer contains a network of thesauri, thus coupling corresponding standardized terms and enabling the intelligent news archive to find corresponding articles in news archives over different countries and languages. This allows journalists and researchers to investigate material on specific topics in a multilingual environment, relying on high result quality and content relevance [5].

RDF (RESOURCE DESCRIPTION FRAMEWORK

The Resource Description Framework (RDF) includes a model to express semantics [3].
The RDF model is simply a model of triples, what makes it very powerful, but difficult to implement. By definition, the RDF descriptions using the triples, the graph or the RDF/XML syntax are equivalent. The RDF/XML parsers read, and verify the RDF/XML syntax, and transform the code written in the RDF/XML syntax in a group of triples and, eventually, in a RDF graph.
RDF is divided in two parts, including two different specifications:

(1) The RDF Model and Syntax Specification (RDFMSS) [, 1999 #26]is a recommendation of the W3C that contains a model to represent RDF metadata, as well as a syntax for expressing and transporting this metadata in a way that maximizes the interoperability of independently developed web servers and clients.

(2) The RDF Schema Specification [6] makes it possible to drawn and to implement in a consistent way, specific metadata vocabularies. These can still be further developed by other projects generating a network of metadata schemas. For example, certain terms of a vocabulary that need to be defined can be specified as refinements of DC [7] elements or of any other previously defined vocabulary. It was already said above that in our metadata description, we used some normalized vocabularies, like DCMES [8] and DC and also some

(News) Text Format Standards like NITF [9] and NewsML [10]. Dublin Core is currently being used in many places, and is one of the foundations that RDF is building on.
The RDF/XML code is not simple. The RDF model is itself of an extraordinary simplicity and the descriptions are of a logical mathematical rigour. However, some parts of the model and of RDF/XML syntax are being currently subject to change. In fact, the W3C "RDF CORE" working group [11], is working in a new specification for RDF with two main goals: (1) to make the specification easier to read and to understand and (2) to remove some assumed mistakes from RDF/XML syntax.
On the other hand, the expression of metadata vocabularies in RDF has also not been a pacific issue for the last years. In 1999 DCMI released a working draft for establishing some rules on expressing Dublin Core in RDF. Since then a great discussion has been made around this subject. The "Expressing Qualified Dublin Core in RDF/XML" [12]working draft released in 2001 covers the most important issues regarding this subject. Although this document is not yet a recommendation from DCMI [13], it already has become a candidate recommendation, which gives us some confidence for using it as a reference guide for our RDF/XML descriptions, in Omnipaper's project context.

OMNIPAPER APPLICATION PROFILE

The concept of application profiles has emerged in discussions on metadata schemas, in relation to work that is being done on metadata registries, specifically in the Dublin Core Metadata Initiative. The application profiles grew out of UKOLN's [14]work on the DESIRE [15] project.

Heery and Patel define application profiles as schemas which consist of data elements drawn from one or more namespace schemas, combined together by implementers and optimised for a particular local application [16]. They state that the principal characteristics of an application profile are that: it may draw on one or more existing namespaces; does not introduce new data elements; it can specify permitted schemes and values; and it can refine standard elements.

In the context of the Omnipaper's project, the definition of the application profile is intended to describe not only the elements from different vocabularies but also elements from our particular concept of application, used in the description of our articles. The definition of all the elements used in the implementation of the RDF/XML metadata structure is fully described, by Yaginuma et al. [17].

The definition of all the elements used in the description of our articles were based on an elaborated analysis of several XML (eXtensible Markup Language) applications like XML NITF (News Industry Text Format) [9], NewsML (News Agency Implementation Guidelines) [10] developed by IPTC (International Press Telecommunications Council) [18], by metadata vocabularies like DCMES (Dublin Core Metadata Element Set) [8], and DCQ (Dublin Core Qualifiers) [7] and by the vocabulary SMES [19]; a set of metadata elements that best define the features of the newspaper articles and other elements were selected. Theses elements make part of the Omnipaper application profile that is illustrated in the picture below and the RDF/XML implementation is shown in [20]. It includes the following six vocabularies:

- Dublin Core Metadata Element Set (DCMES) – http://purl.org/dc/elements/1.1/;
- Dublin Core Qualifiers (DCQ) - http://purl.org/dc/elements/1.1/;
- News Industry Text Format (NITF) - urn:nitf:iptc.org:20010419:NITF;

- News Markup Language (NewsML) - urn:newsml:iptc.org:20010421:NEWSML;
- Omnipaper RDF Schema (omni) - http://www.dsi.uminho.pt/omn/schemas/omn-schema#; and
- vCard - http://www.w3.org/2001/vcard-rdf/3.0#.

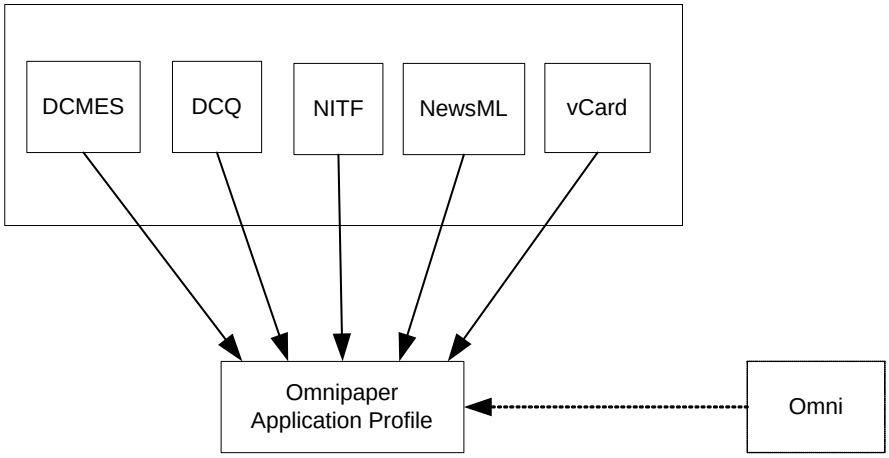

FIGURE 2 - METADATA VOCABULARIES FOR OMNIPAPER

OMNIPAPER SCHEMA

The Resource Description Framework Schema Specification (RDFSS) [3] is a proposed recommendation of the World Wide Web Consortium (W3C) [21]. The schema specification language is a declarative representation language influenced by ideas from knowledge representation (e.g. semantic nets, frames, predicate logic) as well as database schema specification languages and graph data models [3].

The main goal of the RDFSS consists of the definition of a schema specification language that allows the definition of mechanisms needed to define elements from specific vocabularies, to define classes of resources which those elements may be used with, according to restriction mechanisms, based on the constraint rdfs:domain, and in the definition of values for the properties through the rdfs:range.

The RDF Schema specification provides mechanisms of extensibility of RDF vocabularies, through the use of classes and specific properties of the RDF Schema. We can define our own vocabularies that are not more than specializations of vocabularies normalized as it is the case of DCMES (Dublin Core Metadata Element Set).

The concept of RDF schema is directly related with the concept of application profile. While in RDF Schema, is defined a vocabulary that it can be used in the context of one or more RDF applications. In each application profile, it is possible to identify the RDF schema, the elements of each vocabulary and its context in certain RDF application.

Heery and Patel [16] distinguish namespace schema from application profile schema.
Namespace is defined within the W3C XML schema activity [22] and allows for unique identification of elements, which may be defined through an RDF SCHEMA. Within the W3C XML and RDF schema specifications, namespaces are the domain names associated

with elements which, along with the individual element name, produce a URL that uniquely identifies the element [16].

According to Heery and Patel [16] with ‘namespace’ we can

- Identify the management authority for an element set;
- Support definition of unique identifiers for elements;
- Uniquely define particular data element sets or vocabularies.

While in application profiles we can [16]:

- Use elements of one or more namespaces;
- Not introduce new elements;
- Specify schemas and allowed values;
- Refine normalized definitions.

In the Omnipaper project we use some metadata elements that were not previously defined neither in standard vocabularies nor in widely-used vocabularies. For this purpose a specific Omnipaper vocabulary containing these metadata elements was created using RDF schema.
In Figure 2, illustrated below, we show the properties defined in the Omnipaper RDF Schema [23].

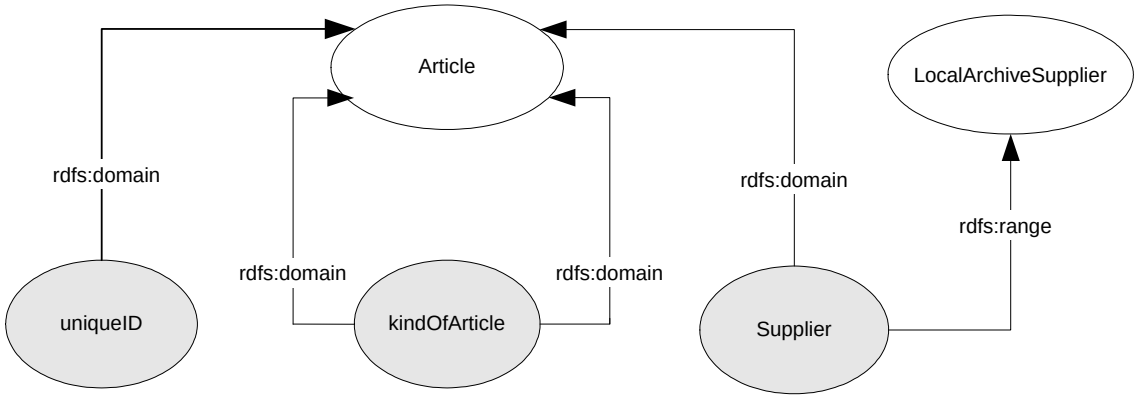

FIGURE 3 - SCHEMA PROPERTIES

uniqueID: Applies to all kinds of articles
kindOfArticle: Applies to all kinds of articles, many has a value any kind of article (any or its subclasses).
supplier: Applies to all kind of local archive owners. May have a value any kind of Local Archive Supplier (any of its subclasses).

Each property has a specific meaning, defines its permitted values, the types of resources it can describe, and its relationships with other properties [, 1999 #26].
In order to support this properties’ definition, we defined in the RDF schema a set of classes as explained below.

The constraint rdfs:range indicates the classes that the values of a property must be members of. Each property can only have at maximum one rdfs:range constraint. “Although it is possible to express two or more range constraints on a property, a similar outcome can achieved by defining a common super class for any classes that represent appropriate values for some property” [6]. For this purpose were defined the classes Article and LocalArchivesSupplier where the classes Interview, Review, OpinionLetter and News were defined as subclasses of the class Article, and the classes PTE, MyNews and Mediargus were defined as subclasses of the class LocalArchivesSupplier, as illustrated in the Figure 3 and 4.

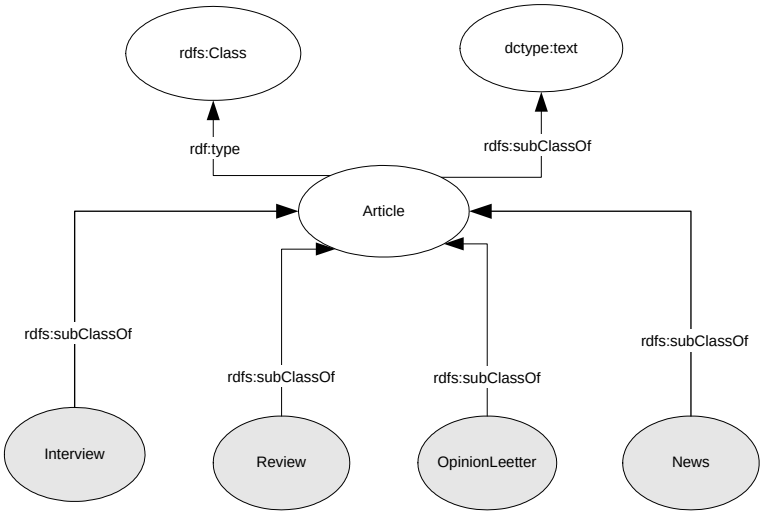

FIGURE 4 - OMNIPAPER SCHEMA: RELATIONSHIP BETWEEN ARTICLE CLASSES

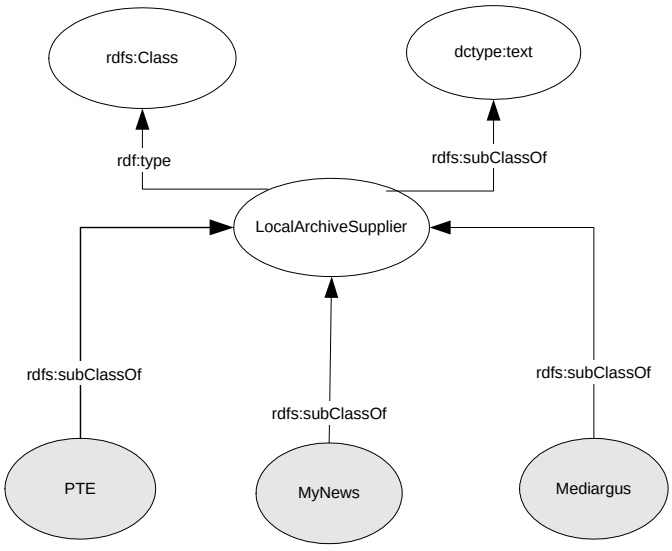

FIGURE 5 - OMNIPAPER SCHEMA – RELATIONSHIP BETWEEN LOCALARCHIEVESUPPLIER CLASSES

METADATABASE IMPLEMENTATION

The prototype for handling RDF layer in the Local Knowledge Layer was accomplished using RDF Gateway [21]. The RDF prototype was developed over RDF Gateway, 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.

The application was developed in a scripting language called RDFQL. RDFQL is based on ECMA Script (Java Script) with query extensions to perform federated searches across multiple data sources. In the RDFQL server pages (RSP) we wrote and deployed on the RDF Gateway the complete definition of our application. The code inside RSP files (RDF Server Pages) interacts with the users and database engine, executing queries within the metadata database and sending answers to the users. Articles that match the query will be displayed by showing a selection of their metadata (title, date and author). By clicking the article, a SOAP request should be sent to retrieve the full text of the article from the news archive server.

The final result is an application that can be accessed using a Web browser by specifying an URL In a first phase of the prototype, we developed the description of the digital articles followed by the implementation of the manipulation process in RDF Gateway, and we also developed a program to transform and upload information to the metadata base. The program transforms article and keyword files into RDF Files and uploads it. All this process off the system architecture is illustrated in the Figure 6.

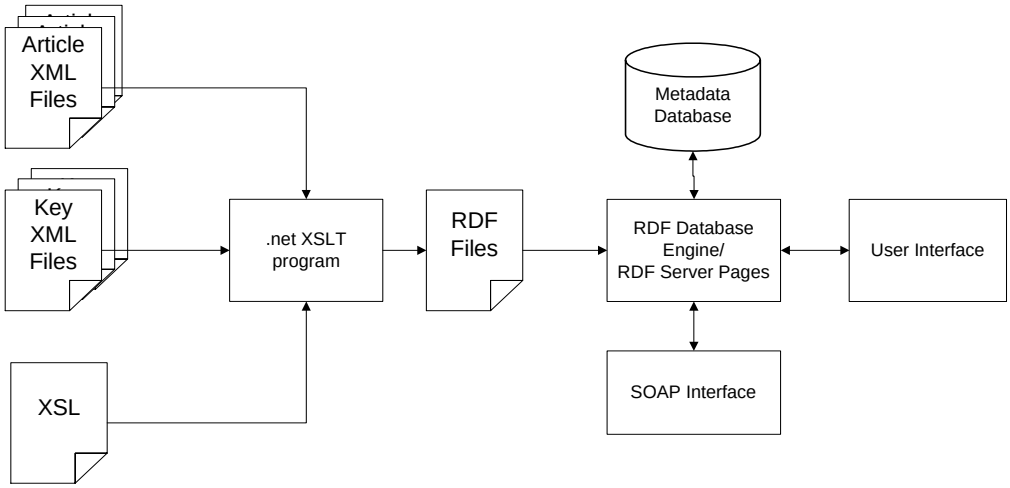

FIGURA 6: SYSTEM DESIGN

CONCLUSIONS AND FUTURE WORK

In the Omnipaper prototype, 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.

In this paper we discussed the development of RDF prototype, in the context of the European digital newspaper articles in the scope of Omnipaper's project.

The RDF prototype tries to investigate efficient ways to describe and store metadata used in the description of the digital newspaper articles and at the same time analyse de efficiency of the RDF technology in the information retrieval provided from different types of distributed information resources.

The metadata approach developed using Resource Description Framework (RDF) and Topic Maps (TM) technologies are being developed in simultaneously and will be cross-tested on a technical level allowing conclusions on both levels.
The cross-testing of the prototypes and comparisons to both metadata structures approach (implemented in Topic Maps and RDF), will allow the project partners to make well considered decisions on the development of the final prototype.

In the next phase we will implement the whole RDF and Topic Maps's prototypes, and then we will accomplish ranking tests, comparing both approaches, according to the comparison criteria previously defined.

NOTES AND REFERENCES

1. *Resource Description Framework (RDF) Model and Syntax Specification.* Feb. 22, 1999. W3C. Available from: <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>
2. *Omnipaper - Smart Access to European Newspapers, EU project IST 2001-32174.* 2002. Available from: <http://www.omnipaper.org>
3. *Resource Description Framework (RDF).* 2003. W3C. Available from:<http://www.w3.org/RDF/>
4. Group, M.o.t.T.O.A.. *XML Topic Maps (XTM) 1.0*. 2001. XTM TopicMaps.org. Available from:<http://www.topicmaps.org/xtm/1.0/>
5. Paepen, B.E., Jan; Schranz, Markus; Tscheligi, Manfred. *Omnipaper: Bringing Electronic News Publishing to a next Level Using XML and Artificial Inteligence*. Proceedings of the 6th International ICCC/IFIP Conference on Electronic Publishing held in Karlovy Vary, Czeck Republic, November 2002.
6. *RDF Vocabulary Description Language 1.0: RDF Schema*. Jan. 23, 2003. W3C. Available from: <http://www.w3.org/TR/rdf-schema/>
7. *Dublin Core Qualifiers*. 2002. Available from: <http://dublincore.org/documents/dcmes-qualifiers/>
8. *Dublin Core Metadata Element Set, Version 1.1: Reference Description.* Feb 4, 2003. Available from: <http://www.dublincore.org/documents/dces/>
9. *NITF - News Industry Text Format*. IPTC. 2002. Available from: <http://www.nitf.org/>
10. Welcome to the NewsML™ Website. IPTC. 2002. Available from: <http://www.newsml.org/>
11. *RDFCore Working Group*. W3C. 2003. Available from: <http://www.w3.org/2001/sw/RDFCore/>
12. Schwänzl, R.; K., Stefan, *Expressing Qualified Dublin Core in RDF/XML*. Apr. 14, 2002. Available from: <http://www.dublincore.org/documents/2002/04/14/dcq-rdf-xml/>
13. *Dublin Core Metadata Initiative*. 2003. Available from: <http://www.dublincore.org
14. *UKOLN* . 2002. Available from: <http://www.ukoln.ac.uk>
15. *DESIRE: Development of a European Service for Information on Research and Education. EU project.*2003. Available from: <http://www.lub.lu.se/desire/>
16. Heery, R. ; P., Manjula . *Application Profiles: Mixing and Matching Metadata Schemas*. Ariadne. Sep. 25, 2002. Available from: <http://www.ariadne.ac.uk/issued25/app-profiles/>
17. Yaginuma, T.P., Teresa; Baptista, Ana Alice, *Metadata elements for digital news resource description*. 2002.
18. *Information Technology for News.* 2000, 2001. Available from: <http://www.iptc.org>
19. SCHEMA, U., *Forum for Metadata Schema Implementors*. 2002. Available from: <http://www.schemas-forum.org>
20. Pereira, T., B., Ana Alice, *Omnipaper Apllication Profile*. 2002. Available from: <http://www2.dsi.uminho.pt/tpereira/OmnipaperApplicationProfile1.rdf>
21. Chappell, G.R., Derrish; Michal , Aaron, *Intellidimension Gateway*. 2002. Available from: <http://www.intellidimension.com/>
22. *Extensible Markup Language (XML) Activity Statement*. W3C. 2002. Available from: <http://www.w3.org/XML/Activity>
23. Pereira, T.B., Ana Alice, *Omniapaper Schema Namespace*. 2002. Available from: <http://www2.dsi.uminho.pt/tpereira/Omni_Schema.rdf>