Searching and Browsing Using RDF-Encoded Metadata: The Case of Omnipaper

Ana Alice Baptista
University of Minho

Abstract: Often metadata applications concentrate on either searching or browsing. Two completely independent prototypes covering both searching and conceptual browsing (with overlapping functionality) have been built as part of the Omnipaper project: one using RDF and one using the Topic Maps family of technologies. This paper will present the full RDF prototype, with particular emphasis on the homogeneity provided by the use of RDF (together with RDF-S) in both approaches. The Omnipaper Resource Description Framework (RDF) prototype handles both these approaches and builds a coherent semantic metadata layer that may be used as a basis for all other prototypes of the Omnipaper project. As a future phase of the project, the main features to be added to the project’s final prototype will be multilinguality and personalization.

Résumé : Souvent les applications pour les métadonnées portent soit sur la recherche, soit sur la navigation. Dans le cadre du projet Omnipaper, deux prototypes complètement indépendants l’un de l’autre, portant à la fois sur la recherche et la navigation conceptuelle (avec la possibilité de chevauchements), ont été réalisés : l’un utilise le système RDF tandis que l’autre utilise la famille de technologies Topic Maps. Cet article présentera le prototype RDF intégral, en mettant un accent particulier sur la pertinence pour les deux actions du système RDF (de pair avec RDF-S). Le prototype Omnipaper qui emploie le système RDF peut faciliter les deux actions et développer un niveau sémantique de métadonnées qui soit cohérent et qui puisse servir de base pour tous les autres prototypes du projet Omnipaper. Dans une étape future, les caractéristiques principales qui viendront s’ajouter au prototype final du projet sont le multilinguisme et la personnalisation.

Keywords: Metadata; Omnipaper project; Internet searching, browsing

The Dublin Core Glossary defines metadata as “information that expresses the intellectual content, intellectual property and/or instantiation characteristics of an information resource” (Woodley, 2001). In this article, a metadata application (MA) is considered to be an application that handles metadata, regardless of its goal.
Often metadata applications concentrate on either searching or browsing. When the application is designed for searching, catalogues of information resources are built and are then searched or indexed for searching. Browsing across these catalogues is often achieved through the explicit and implicit relationships between them (for example, relationships such as references, multiple-versioning, and so on). One of the basic motivations for building these catalogues is to facilitate resource discovery, making this process more efficient and effective across the Web. In this paper, this approach is called the “Low-Level Approach” (LLA).

On the other hand, when the application is built primarily for browsing, a network (or web) of concepts is built, based on some kind of knowledge organization (ontology). In this approach, the main motivation is the desire to be able to browse through a network of concepts that link to resources, with these links having specific meanings. In this paper, this approach is called the “High-Level Approach” (HLA).

Each of these approaches typically uses a different set of technologies that best reflects its philosophy. The Low-Level Approach usually uses HyperText Markup Language (HTML) (http://www.w3.org/HTML), plain Extensible Markup Language (XML) (http://www.w3.org/XML) or the Resource Description Framework (RDF) (W3C, 1999). The High-Level Approach typically uses Topic Maps ™ and its XML application, XML Topic Maps, or XTM (TopicMaps.Org Authoring Group, 2001), or RDF Schema (RDF-S) (W3C, 2000), or other ontology languages based on RDF-S, such as DAML+OIL (DARPA, n.d.; Horrocks et al., 2001) and OWL (W3C, 2003).

The worldwide metadata community is often divided by these approaches and their underlying technologies. Metadata applications usually handle one of these approaches, but it is not very common that they handle both.

There is also division over the use of the RDF-S or TM families of technologies for the High-Level Approach. Each of these technologies has its own community of developers and implementers, and although they are competitive technologies in some sense, they also complement each other. Efforts for describing TM using RDF and vice-versa have been made in the past (Dumbill, 2001; Lacher & Decker, 2001; Ogievetsky, 2001).

In most cases where an RDF application handles the LLA, some of the metadata elements refer to words or codes from existing controlled vocabularies, but they are only used for searching the catalogues. Usually they do not allow users to navigate through those vocabularies and find information the other way around.

Similarly, in most cases where an RDF application handles the HLA, while its records may point to real resources, it does not have searchable catalogue information about those resources. These applications are most often targeted to high conceptual navigation or browsing, not to providing search functionality.

There are some applications that handle both LLA and HLA approaches (for instance, MPress: http://mathnet.preprints.org). The Omnipaper RDF prototype goes one step further by implementing not only a subject thesaurus but also a
lexical thesaurus on the HLA; these are directly connected with the articles’ descriptions in the RDF metadatabase. These new developments allow the addition of important functionality, including manual and automatic query expansion.

The Omnipaper project
The Omnipaper (also known as “Smart Access to European Newspapers, IST-2001-32174”) project is a European project that 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; and (4) lift widely distributed digital collections to a higher level. The Omnipaper consortium has eight partners from four European countries, three of them being digital news providers.

Its prototype architecture is based on potentially distributed metadata databases (“metadatabases”) that provide information to a centralized upper layer, the main goal of which is cross-archive navigation and searching. This layer is then enriched with multilingual and personalization functionality for interface with the user (see Figure 1).

The metadata view of the system consists of two layers: the Local Knowledge Layer (LKL) and the Overall Knowledge Layer (OKL), as shown in Figure 1. The LKL is composed of distributed metadatabases, one for each local archive of digital news. Thus, in accordance with the terminology used in this paper, it implements the LLA.

The OKL consists of an ontological metadatabase, the purpose of which is to facilitate cross-archive browsing and navigation using a web of concepts. Thus, in accordance with the terminology used in this paper, it implements the HLA.

The original plan was for the LKL to be implemented using RDF and the OKL using TM, then to integrate both prototypes. However, at the beginning of the project, the participants agreed instead to develop two parallel prototypes, one using RDF and the other using TM, then to test each part of the two prototypes against the other. The parts with the best performance would then be integrated into a final combined prototype. The following sections present the overall solution for the RDF prototype.

The Omnipaper RDF prototype
The development of the Omnipaper RDF prototype comprised the following general steps:

1. definition and development of the metadatabase: the LLA;
2. definition and development of the conceptual layer (subject + lexical thesaurus): the HLA; and,
3. integration of previously developed prototypes into one full prototype.
Implementing the Low-Level Approach—The Local Knowledge Layer

The first phase involved the conceptualization of the project. We looked at what was being done with news articles with regard to metadata and found that all the projects still focused primarily on article content, although metadata was also being included (Yaginuma, Pereira, & Baptista, 2003).

As a consequence, the Omnipaper team needed to find and agree on a set of metadata elements (the Omnipaper application profile) to be used for the project. The selection of the exact elements to include in this application profile was made very carefully, because they constituted the underlying data upon which all functionality was to be built. Several metadata-element sets were carefully analyzed in terms of the semantics of their elements, their usage in the community, and their interoperability across communities (Yaginuma et al., 2003).

For the Omnipaper application profile, we chose to use the Dublin Core Metadata Element Set (DCMES) and Dublin Core Element Refinement Qualifiers (DCERQ) as the basic set of metadata elements, to which elements from other vocabularies (pre-existing or created by us) would be attached. The choice of DCMES was a natural one for a number of reasons:
1. It has been a stable element set since 1996.
2. It has been a Dublin Core Metadata Initiative (DCMI) (http://dublincore.org) recommendation since 1998 with its version 1.0 (DCMI, 1998) and since 1999 with its version 1.1 (DCMI, 1999).
4. It has been adopted by CEN/ISSS.
5. It has an official positioning within the World Wide Web Consortium (W3C) (http://www.w3.org).
7. And, above all, it is a widely used metadata-element set all over the world.

DCERQ has been used mainly because it is also part of a DCMI recommendation and complements well DCMES for general resources’ description.

As shown in Figure 2, the application profile (Heery & Patel, 2000) was then enriched with some elements from vCard (Dawson & Howes, 1998; Iannella, 2001) and some new elements created by us, which were inspired by the News Industry Text Format (NITF) (http://www.nitf.org/) and News Markup Language (NewsML) (http://www.newsml.org/pages/index.php) XML document type definitions (DTDs) or schemas, and our own needs. In the end, the application profile had a set of 27 elements divided into six categories: Article Identification, Article Ownership, Article Location (Storage), Article Relevance/Audience, Article Classification, and Link Information (Yaginuma et al., 2003).

**Figure 2: Contributions to the definition of the Omnipaper Application Profile**
A new namespace RDF Schema had to be created for defining all the new metadata elements we wanted to include in the application profile. Together with these elements (RDF properties), some classes were created in order to be able to specify usage contexts for them, that is, allowable values for those properties (Pereira, Yaginuma, & Baptista, 2003b).

With all these steps completed, rules for metadata encoding were established and a template was built. We took into account all the recommendations made by Kokkelink and Schwänzl (Kokkelink & Schwänzl, 2001), although this document was still a proposed recommendation (DCMI) from the DCMI. From the templates developed in RDF/XML (Pereira & Baptista, 2003; Pereira, Yaginuma, & Baptista, 2003a), we created the corresponding XML Schemas that would be used to validate our RDF/XML files.

After validation, these RDF/XML files are uploaded to the metadatabase and converted to RDF triples. As the metadatabase platform, we chose to use RDF Gateway®, a Microsoft® Windows™-based native RDF database management system combined with a HTTP server. Some RDF Server Pages (RSPs) were created in order to provide some functionality for the end user. The overall architecture for this layer can be seen in Figure 3.

As this prototype was meant to implement the LLA, only a searching functionality was added. Functionality for navigating documents’ metadata could have been implemented (by making use of relational metadata elements that hold uni-
versal resource identifiers to other resources), but it was not. Figure 4 shows a printscreen of the user interface built for the first Omnipaper RDF prototype.

![Printscreen of the first RDF prototype for the Local Knowledge Layer](image)

Implementing the High-Level Approach—The Overall Knowledge Layer
As with the LKL, the approach to the OKL prototype also began with a conceptualization phase. The goal was to have an ontology-based web of concepts that would be, in some way, linked to the articles themselves.

The news-provider Omnipaper participants had worked with different ontologies. Some of them had limited themselves to categorizations according to newspaper sections. These categorizations are highly subjective (for example, what is considered to be local news on a European scale may be considered national news on a country scale) and very general (sports and national politics, for example, are broad categories), and as such, they were not suitable for the level of refinement we wanted to use to empower our system. Some other ontologies our partners proposed were very complex and had not gained acceptance for use by other systems. The solution agreed upon was to use the International Press Telecommunications Council Subject Reference System (IPTC SRS) (IPTC, 2003).

The IPTC SRS is organized hierarchically and has no associative (thesaurus-like) relationships between its concepts; as such, it is not very rich semantically.
However, it is sufficient for providing a high-level navigation layer, thus implementing the aforementioned HLA.

Several RDF-based vocabulary and ontology description languages were studied in order to choose one to codify the IPTC SRS (Pereira & Baptista, 2004). In the end, RDF Schema (W3C, 2000), an RDF vocabulary-description language, was chosen, mainly due to the fact that the IPTC SRS vocabulary is so simple that a more powerful (and complicated) language would be useless and inappropriate.

The connection between the OKL and the LKL is made through the “dc:subject” metadata element as shown in Figure 5; that is, only values from the IPTC SRS can be used in the dc:subject metadata element for each description stored in the metadatabase. Conversely, dc:subject-based indexes may be stored together with each concept (value) of the IPTC SRS. For performance reasons, this may turn out to be an important feature in later phases that use a much larger metadatabase.

To add value to the OKL navigation and searching functionality, another empowering information-organization tool was included and linked to the metadatabase: WordNet®. WordNet is “an online 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” (Princeton University Cognitive Science Laboratory, n.d., p. 1).

WordNet version 1.6 was also implemented in RDF both using RDF and RDF-S technologies. It was downloaded and included in a local metadatabase, and its connection to the articles metadatabase was made through the “omni: key_list” metadata element as shown in Figure 5. This is a very natural feature that can be used to perform query expansion, whether it is done automatically or manually.

So far no direct relationship between WordNet and the IPTC SRS has been implemented. The only relationship is that when a search for a concept is performed in the IPTC SRS, the same search is performed in WordNet for that particular word. It would be of major interest to implement word clustering into IPTC SRS concepts, but it is beyond the scope of the Omnipaper project for the time being.

More functionality, particularly navigation functionality, was added to this prototype. This functionality implements the HLA by defining IPTC SRS concepts as a type of dc:subject metadata-element content. This allows users to navigate through the IPTC SRS tree (the left side of the screen on Figure 6) to find articles that have been indexed according to the subjects listed in the tree. Furthermore, because WordNet is also part of the system, when a user navigates the IPTC SRS, related words (associated with the branches of the IPTC SRS) appear on the screen (see the top-middle section of Figure 6). The user can click on these words (synonyms, antonyms, et cetera) to access results not previously retrieved by the system. When the user clicks one of these words, the system simply expands the previous query to search for that word in the omni: key_list metadata element’s contents.
Conclusions and future work

In the course of the Omnipaper project (Smart Access to European Newspapers, IST-2001-32174), it was decided that two completely independent metadata prototypes would be developed and compared: one based on RDF and the other on TM technologies. This article described the main features and decisions made through the development of the RDF prototype, both for the Local Knowledge Layer and for the Overall Knowledge Layer.

It is very common to see RDF applications that concentrate on either searching or browsing, but it is more rare to find RDF applications that are designed and built to support both approaches. It is even less common that these applications combine a subject thesaurus with a lexical thesaurus to enhance both user queries and navigation.

In the first phase, an RDF metadatabase was created to handle the news-article descriptions (implementing the Local Knowledge Layer); then a news-specific thesaurus was encoded using RDF-S, to which an RDF-encoded lexical thesaurus was added (implementing the Overall Knowledge Layer). These two layers have been linked through the Subject and Key_list metadata elements.
The RDF technology proved once again to be well suited to both the metadata Low-Level Approach and the metadata High-Level Approach, and the project demonstrated that the relationship between these approaches is a natural one in RDF terms.

It is important that future work emphasize the comparison between the RDF and Topic Maps prototypes. The final decision was to integrate them in a way that could take advantage of the best features of each prototype. The associated tests, results, and discussion will be published in a future article.

At the current time, multilinguality and personalization functionality is being added to the final integrated prototype.

Acknowledgments
I would like to thank Roland Schwänzl and João Álvaro Carvalho for having kindly reviewed this article at my request.

References
DARPA. DAML—The DARPA Agent Markup Language Homepage. URL: http://www.daml.org [March 6, 2004].


