METADATA ELEMENTS FOR DIGITAL NEWS RESOURCE DESCRIPTION

Tomoko Yaginuma, Teresa Susana Mendes Pereira, Ana Alice Baptista

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

RESUMO

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 (Campbell, 2002), internet resources retrieval (Neuroth, et al, 2001) and agricultural resource cataloguing (Onyancha, et al, 2001). The IST-funded OmniPaper project 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). 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 (Schranz, et al, 2002):

(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)[1]. This means that each local news provider does not need to make any change on its text format.

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) (Lassila; et al, 1999) and XML-Topic Maps (XTM) (Pepper, et al, 2001) 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 two global news format standards widely used in the news industry, that are NITF and NewsML[2]. Both are implemented and maintained by the International Press

[1] http://www.nitf.org/
[2] http://www.newsml.org/

Telecommunications Council (IPTC), that is an organisation developing and publishing industry standards for the interchange of news data.

NITF is a text format standard which defines the content and structure of news articles (IPTC, 2002). 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 throughout the news content.

NewsML is a one level higher standard for news representation and management (IPTC, 2002). 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.

Although these standards are not specifically for metadata description but for news text formatting, they have a variety of metadata elements within the structures. As the efficient use of metadata is one of the most important features of the standards, we studied their use 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 NewML 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 NewML's metadata. Some of the metadata is well designed to be able to add or replace the 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.

Considering all the above, it might be beneficial to use NITF or NewsML for the text format for describing the OmniPaper participating 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 it for dealing with metadata in the following context.

The documents written in these standards contain not only metadata information but also the entire news content, 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. Besides, metadata elements spread all over the document. There is no one container to gather all the metadata elements within the text structure. 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.

As a result, OmniPaper does not use NITF or NewsML for metadata handling although they were largely studied as a reference. Instead, our choice for metadata description went to Dublin Core Metadata Element Set (DCMES) 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 (DCMI, 2002). 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 (Berners-Lee, et al, 2001).

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 or NewsML. The last option is OmniPaper-original metadata, that is the creation of elements. This must be avoided unless there is strong necessity.

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 DCQ, 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 Dublin Core Qualifiers (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.

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

Table 2 - Article Ownership.

Name	**Description**	**Source**	**Encoding Scheme(s)**
Copyright	Container for copyright information.	NITF	
Owner	The local archive that owns the article.		vCard

Table 3 - Article Location (Storage).

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

Table 4 - Article Relevance/Audience.

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

Table 5 - Article Classification.

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

Table 6 - Link Information.

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

It might be necessary for certain elements to state additional explanation, the reason why it is selected and so on.

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.

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 (Pereira, et al, 2003).

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: KindOfText; Section; Copyright; OfInterestTo; Spatial.

The element KindOfText 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, KindOfText 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 (DCMI, 2002), 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 (DCMI, 2002). 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 two existing news text format standards, which are NITF and NewsML, 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. 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] Berners-Lee, Tim; Hendler, James; Lassila, Ora, *'The Semantic Web'* http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2, 2001.

[2] 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.

[3] DCMI Education Working Group, http://dublincore.org/groups/education/

[4] DCMI Type Vocabulary, http://dublincore.org/documents/dcmi-type-vocabulary/

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

[6] Dublin Core Metadata Initiative, http://dublincore.org/

[7] Dublin Core Qualifiers, http://dublincore.org/documents/dcmes-qualifiers/

[8] International Press Telecommunications Council, http://www.iptc.org/

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

[10] 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.

[11] News Industry Text Format, http://www.nitf.org/

[12] News ML, http://www.newsml.org/

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

[14] 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.

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

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