Library and Information Services in Astronomy IV
July 2-5, 2002, Prague, Czech Republic
B. Corbin, E. Bryson, and M. Wolf (eds)

Investigation of Metadata Applications at Palermo Astronomical Observatory

Francesca Martines & Flavio Morale
INAF - Osservatorio Astronomico di Palermo "G.S. Vaiana", Piazza del Parlamento, 1, 90134 Palermo, Italy
martines@oapa.astropa.unipa.it & morale@oapa.astropa.unipa.it.

Abstract. In this paper we report our investigation of metadata applications at the Palermo Astronomical Observatory. We provide a brief introduction to metadata, followed by an explanation of needs and goals. We describe instruments and tools we have selected and tested, and we give the results we have obtained. We focus especially on our attempt to develop rules for the creation and management of a particular type of internal technical report. Possible applications, along with open issues and proposed solutions, are also provided.

1. Introduction

Metadata is a very popular word in information technology today. If you search on the Web with a search engine for "Metadata", you can be sure that you will find about a million and half of occurrences of the term. It seems that almost everybody in the information community has to face this concept. But what is metadata? And - more important - why is it so relevant?

Metadata is usually defined as "data about data" and apparently there is a slight difference in meaning, depending on context.

A typical example of metadata is the traditional library card catalogue: each card gives information about a resource (the book), using elements which are "physically" included in the resource itself (author, title, publisher, etc.), plus other details which are "added" by librarians (subject, classification code, call number, etc.).

It is not obvious to distinguish data from metadata, since it often depends on one's point of view. The underlying idea is that a standardised way of describing resources will produce a (more) effective retrieval of information. Moreover, standardised descriptions could facilitate interoperability between different systems and platforms. There are two main approaches to interoperability:

- cross system search: there is no sharing of data/metadata but search engines are mapped to a common set of search attributes. This is, for instance, the Z39.50 protocol philosophy.

- metadata harvesting: it basically consists in a migration from original metadata to a common set of elements, which are then harvested. This is, for instance, the Open Access Initiative approach.

It is also possible to decide if metadata and related documents/objects have to be stored together, in the same location (embedded metadata), or separately. It depends on the format of document/object and on the language used to mark it up. Each one of these solutions has pros and cons.

2. The Investigation

We have started our investigation about metadata applications, looking for a solution to the following necessities, which are partially complementary:

- To create (and to retrieve) electronic documents which include data and metadata in a standardised format that provides automatic extraction of metadata (i.e. information about resources/documents)

- To create standardised metadata for existing documents, in different formats, in order to make these documents and/or archives interoperable.

To obtain these results, we need to select the following elements:

1. Markup Language
2. DTD (s)
3. Metadata Description Standard
4. Validating parser(s)
5. Tools for output format(s) of files
6. Tools for automatic extraction of metadata

3. Instruments/Tools

3.1. The Markup Language

In order to choose the more suitable syntax, we looked at semantic languages, such as the powerful SGML (Standard Generalized Markup Language). Instead,we decided to use XML (eXtensible Markup Language), a semantic language derived from SGML, that prescribes how to embed descriptive mark-up language within a document in a standard format.

3.2. The DTD

An XML document refers to a DTD (Document Type Declaration), which describes the logical structure of the document and its parameters. Subsequently, a parser can test a document, checking if it is well-structured (i.e. grammatically and logically correct) and also, more important, if it is valid, (i.e. conforms to the DTD). We chose DocBook, which is a very popular set of tags for describing books, articles, and other prose documents, particularly technical documentation, for its capability to structure complex documents. DocBook Elements also include elements that can be used for both metadata and normal extraction.

3.3. The Metadata Standard

Apart from the language, a standard for describing resources was requested; and, we have decided to adopt Dublin Core Metadata Element Set, for its wide diffusion and acceptance, which is a guarantee for its development and maintenance. The DCMES can be used for both standardisation in creating metadata, its subsequent extraction and for mapping search elements. Each of its 15 elements (see below), is optional and repeatable and can be specified by special and determined qualifiers.

3.4. The validating Parser

To validate our XML documents we have successfully used both xmllint, a parser coming with the libxml distribution and (o)nsgmls, another validating parser.

3.5. The output format

To produce acceptable and possibly different output formats for XML files, we have used both openjade and xmlto succeeding to have RTF, HTML, TXT and PS output of our files. There is also the possibility to have other formats such as PDF.

3.6. Tools for automatic extraction of metadata

We have still to find the way of extracting metadata automatically from our XML/DocBook files. There are two main options:

- to find and test an already existing tool which produces what we need (hopefully!)
- to set up ourselves a suitable tool, possibly using a XLST macro.

All tools have been installed and tested on a Linux Red Hat 7.2 platform.

4. A Test Case

In a test case, we attempted to create a model for an internal technical report that was related to activities of Observatory EDP staff.

With some modification due to the particular nature of these documents,we sought a standard and decided to adopt the ANSI/NISO Z39.18-1995 Standard (Scientific and Technical Reports Elements, Organisations and Design).

We identified some elements that can be extracted as metadata and we tried to map them into the Dublin Core Metadata Element Set (see table after). If compared with the complete list of DC Elements, the results concluded that almost every element had been mapped, with the exception of DC Coverage.

Then, we tried to map selected DocBook Elements into Dublin Core Elements. This operation was made considering Simple Dublin Core, but we plan to extend the experiment to Qualified DC. There were some problems since DocBooks Elements are more numerous and unique than DC elements. At this stage, our mapping should be considered tentative.

Table 1. Mapping Report Elements into Dublin Core Elements

REPORT (Front Matter)	INTERVENTION (Text/body)	DC Element
Report Number	Operation Number	Identifier
URI of Report		Identifier
Author of Report	Author of operation	Creator
Title of Report	Title of operation	Title
Abstract		Description
Date of report opening	Date of operation starting	Date
Date of report closing	Date of operation ending	Date
Date of report modification		Date
Publisher		Publisher
Warning / Distribution List		Rights
	Operation origin	Relation/Contributor
Keywords		Subject
Document type (report)		Type
Document type (txt)		Format
Language		Language
Revisions/links to other reports		Source/ Relation

5. Possible Applications and Open Issues

There are several possible applications in our Observatory. Here is a small sample list as follows:
- Inclusion of metadata in Observatory Web pages.
- Create a standard for internal reports/documents of general interest (i.e. from administrative department, library, EDP office, etc.).
- Creation of lists of requested/ordered/arrived books or other general items.
- Use in SDI, for instance, creating lists of "New and Forthcoming Books" directly from publishers' electronic newsletters.
- Management of Observatory staff publications, making the systems interoperable with related full text documents (local or remote).

There are also some open issues that need to be solved. The following ones are the most relevant:
- There is still a need devise a method for extracting metadata and convert it to DC metadata.
- Careful investigation is required for Qualified Dublin Core applications.
- Taking into account the ongoing work of the DocBook Technical Committee, DocBook/DC mapping needs to be refined

References

ANSI/NISO Z39.18-1995 Standard: Scientific and Technical Reports Elements, http://www.niso.org/standards/resources/Z39-18-1995.pdf

Dublin Core - http://www.dublincore.org

DocBook - http://www.docbook.org

Hodge, G. - Metadata made simpler - NISO Press, 2001 - http://www.niso.org/news/Metadata_simpler.pdf

nsgmls - http://openjade.sourceforge.net/doc/nsgmls.htm

openjade - http://openjade.sourceforge.net

XML - http://www.w3.org/XML/

xmllint - http://xmlsoft.org

xmlto - http://cyberelk.net/tim/xmlto