Library and Information Services in Astronomy III
ASP Conference Series, Vol. 153, 1998
U. Grothkopf, H. Andernach, S. Stevens-Rayburn, and M. Gomez (eds.)

Metadata: Standards for Retrieving WWW Documents (and Other Digitized and Non-Digitized Resources)

Diann Rusch-Feja

Max Planck Institute for Human Development, Lentzeallee 94 D-14195 Berlin, Germany, e-mail: ruschfeja@mpib-berlin.mpg.de

Abstract. The use of metadata for indexing digitized and non-digitized resources for resource discovery in a networked environment is being increasingly implemented all over the world. Greater precision is achieved using metadata than relying on universal search engines and furthermore, metadata can be used as filtering mechanisms for search results. An overview of various metadata sets is given, followed by a more focussed presentation of Dublin Core Metadata including examples of sub-elements and qualifiers. Especially the use of the Dublin Core Relation element provides connections between the metadata of various related electronic resources, as well as the metadata for physical, non-digitized resources. This facilitates more comprehensive search results without losing precision and brings together different genres of information which would otherwise be only searchable in separate databases. Furthermore, the advantages of Dublin Core Metadata in comparison with library cataloging and the use of universal search engines are discussed briefly, followed by a listing of types of implementation of Dublin Core Metadata.

1. Metadata: Definition and Typology

Metadata are data about other data and objects. Metadata are used to describe digitized and non-digitized resources located in a distributed system in a networked environment. To be effective, they must be standardized. Traditional metadata include library cataloging rules, schemes, and formats. Due to the expanding electronic information environment, these have been expanded to reflect the needs of information discovery and use of information in such a networked environment and thus, rather than "electronic cataloging rules", the term "metadata" has been used. This also reflects a movement away from the traditional focus on books and printed works in libraries to a much more expanded focus on all kinds of data and objects, including digitized objects. Thus, metadata can include bibliographic information such as that in traditional library catalogs, subject cataloging, such as descriptors, classification designations, abstracts, etc., structural data on the type and size of resources, as well as technical requirements for their use or necessary for access, relationships (thematic, formal, references, citations, etc.), terms and conditions for obtaining and using the resources, etc.

In a relatively early article, Bearman & Sochats (1994) defined six types of metadata (see http://www.oclc.org:5046/conferences/metadata/requirements.txt), related to

1. identification and resource discovery purposes ("resource discovery")
2. access conditions and usage requirements ("terms and conditions")
3. structural aspects ("structure")
4. contextual aspects ("context")
5. content aspects ("content")
6. the use of this resource ("use history")

This typology of metadata applies to most of the metadata which could conceivably be used for any digitized or non-digitized resource. Up until now, only the first two sections of the Bearman/Sochats typology have been developed to any degree. The third group of metadata in the Bearman/Sochats typology refers to the structure of the object itself – this may go to the level of arrangement into chapters, sections, data segments, etc. Context metadata give information on the context in which the object being described originated. Content metadata, according to Bearman/Sochats, give a much deeper analysis of the context aspects than the usual context descriptors used for resource discovery (point 1) or in documentational analysis. These and the last type of metadata in the Bearman/Sochats typology have not yet been as developed as the first two – probably because of the fact that these belong to a more sophisticated use of metadata than has thus far been achieved in the networked environment – although working groups are currently developing all areas. In addition to metadata which refer specifically to the object itself, there has also been a need for administrative metadata found by some of the implementers of metadata. These administrative metadata are added locally to distinguish aspects of holdings, individual subscription information, access information, etc.

2. Forms of Metadata and Various Metadata Sets

Metadata are most frequently used in three ways:

1. As <META Tags> in the HEADER of HTML-documents
2. As a separate file of Meta-Information <META =... > to describe a non-HTML-file (sound, image, or program file)
3. As a database category in a subject-oriented WWW-server or distributed information system (with its own Harvester and/or search engine)

In addition to these forms of metadata, based on the wide use of HTML and being accommodated for in HTML Version 4.0, the further development of metadata in RDF ("Resource Description Framework") by the WWW Consortium and the integration of metadata into XML (EXtensible Markup Language; http://www.w3.org/TR/REC-xml) structures enhances possibilities for its use. In view of this, the W3 Consortium are working together with the developers of Dublin Core Metadata to implement the use of Dublin Core Metadata into the XML environment. Meetings to this end are currently being held parallel to this conference at the 7th WWW-Conference in Brisbane, Australia.

There are various sets of metadata, in fact more than are mentioned here. The major metadata formats include:

TEI: Text Encoding Initiative dating to 1988 developed by the Virginia Text Center, USA;

WHO IS++Templates: an early form of template-oriented metadata to describe networked items, originally used for mailing lists, then extended to various other resources;

GILS: Government Information Locator Service (http://www.usgs.gov/gils/) (used by the US and Canadian Governments);

EAD: Encoded Archive Description (http://lcweb.loc.gov/loc/standards/ead/) used by the US Archives;

MARC: Library format in many countries of the world (with slight national distinctions – AUSMARC, UKMARC, DANMARC etc.);

DC: Dublin Core Metadata (which will be focussed on in this paper);

DOI: Digital Object Identifier (which includes the Dublin Core Metadata for describing the object itself, as well as additional publisher-oriented information, such as price and purchase conditions extending down to the level of the individual article, diagram or graphic).

TEI and GILS (as well as EAD, WHOIS ++ Templates and other metadata sets prior or parallel to the development of the Dublin Core) are extremely text-oriented. The DOI (http://www.doi.org/) was only introduced in October 1997 and has been developed by five publishers or publishing aggregates including the American Publishers' Assocation and the American Medical Association. The DOI has – at the date of this presentation still inofficially – adapted the Dublin Core Metadata Set for the bibliographical and subject-oriented description aspects within the DOI, and will also hold a variety of additional subscription, purchase- and use-oriented metadata necessary for the publishers' purposes. The DOI will also point to metadata indicating the place within the production chain (i.e., whether the text is in the peer review process, whether and when it has been accepted, what issue it will be in, etc.). In the following discussion, this paper will focus on the Dublin Core Metadata, since this set of metadata seems to be developing into a standard for describing all types of Internet resources in various subject domains and geograhical regions of the world.

3. Dublin Core Metadata, their Elements and Qualifiers

The Dublin Core Metadata were founded by a group of librarians, information scientists, and other parties interested in describing Internet resources for more precise retrieval than was possible via the universal search engines. They met for the first time in April, 1995, in Dublin, Ohio, (hence the name "Dublin Core", or DC) at the invitation of the OCLC organization and the United States' National Science Foundation. The outcome of this first Dublin Core Metadata Workshop was the designation of 13 elements which the participants agreed upon as necessary to meet the requirements of the metadata described in the first point, "resource discovery", of the Bearman/Sochats typology. In the ensuing four Dublin Core Workshops (http://purl.org/metadata/dublin_core/workshop.html), the 13 Dublin Core elements were refined, the overall "Warwick Framework" for encorporating various metadata sets was developed, and options of extensibility for Dublin Core Elements were delineated. In December 1996, the set was expanded

to 15 elements. The rationale for having only 15 elements was to keep the resource description for purposes of discovery (via search machines and individual Harvester gatherers) to a minimum, although it is recognized that certain detailed and structured information will also be necessary (see below "Qualifiers"). The 15 DC Metadata Elements (http://purl.org/metadata/dublin_core_elements/) are listed here with a short description of each element (more detailed descriptions of the individual elements can be found via the Dublin Core Home Page (http://purl.org/metadata/dublin_core):

DC.Title	Title of the Resource
DC.Creator	Author, Creator
DC.Subject	Subject, Keyword
DC.Description	Annotation, Abstract, etc.
DC.Publisher	Publisher (Person or Institution)
DC.Contributor	Contributing Person or Institution
DC.Date	Date (see separate list of Sub-Elements "DC.Date")
DC.Type	Resource Type (according to a list of accepted terms)
DC.Format	Format, File Type, also Physical Medium
DC.Identifier	Resource Identification: URL, URN, ISBN, etc.
DC.Source	Resource (physical, digital) from which the current resource was derived, digitized, etc.
DC.Language	Language of the Resource
DC.Relation	Relationship to other Works
DC.Coverage	Geographic or Temporal Coverage
DC.Rights	Rights Management Statement (or Link to), Copyright

In the workshop series, two groups of Dublin Core developers crystallized: the minimalists who maintain that the 15 elements are simple and sufficient enough to be used for resource discovery, and the structuralists who maintain that in order to obtain precision in searching sub-elements and other qualifying aspects are necessary. Thus, since the fourth Dublin Core Workshop, three qualifying aspects have been accepted to enable the Dublin Core to function in an international context and also meet higher level scientific and subject-specific resource discovery needs. These three Dublin Core Qualifiers are:

LANG: indicating the language of the contents of the META-information, to be used in both resource discovery and in filtering retrieval results

SCHEME: indicating the set of regulations, standards, conventions or norms from which a term in the content of the META-Information has been taken

SUB-ELEMENT: refinement of some of the DC elements to gain more precision

LANG is important for establishing greater internationality and serving needs of non-English language groups in describing their resources. It not only serves to distinguish specific key words, titles, etc., according to language, but can be used for filtering. For instance, the LANG-field of abstracts for foreign language articles could indicate if an abstract is available in another language. An example of the LANG-Field is given in the following:

<META=DC.Title, CONTENT="Zeitschrift für Pädagogische Psychologie" (LANG=de)>
<META=DC.Title, CONTENT="German Journal of Educational Psychology" (LANG=en)>

The SCHEME-Qualifier can be used in various ways. It allows distinction of several classification schemes, different abbreviation conventions, different

date conventions, etc. The following example shows the differentiation which is possible using the SCHEME-Qualifier:

<META=DC.Title.Alternative, CONTENT=(SCHEME=SCCI Citation Title ISI) "Z PADAGOG P" >

In this example, the alternative title is an abbreviation of the above title used by the Social Sciences Citation Index.

<META=DC.Subject, CONTENT=(SCHEME=DDC) "370.15" >
<META=DC.Subject, CONTENT=(SCHEME=LOC) "LB 1051" >

This example shows how, by using the SCHEME Qualifier, other possibly meaningless or at best indistinguishable numeric or alphanumeric classification numbers can be searched in view of a certain classification system. Using the SCHEME qualifier, notations from various classification systems and thesauri can be used for the same object.

4. Internationality via Metadata

Although in the field of astronomy, internationality has long been achieved, there are still various subject areas which have not achieved this level of internationality. Hence, certain advantages of using metadata for achieving internationality are listed here: Metadata can include various language variations, use language limiters to filter search results, include abstracts in various languages, integrate multilingual thesauri and international classification schemes, provide links to translations, related works, and international cooperation partners, etc.

5. Sub-Elements

As an example of the refinement of an individual element through the use of subject elements in order to gain more precision, the example of DC.Date has been chosen. The diversity and the necessity of having sub-elements for certain Dublin Core Elements (but not for all) should be evident. The DC.Date, when not qualified, is considered to be the date of creation or of first publication on the net. However, a vast number of other dates may be relevant for the object. thus, DC.Date can be qualified in the following sub-elements:

DC.Date.Created	DC.Date.Available
DC.Date.LastModified	DC.Date.Verified
DC.Date.Published	DC.Date.Accepted
DC.Date.Expired	DC.Date.DataGathered

The additional sub-elements provide additional information which may or may not necessarily serve as search criteria, but also as filtering criteria.

6. Relationships to Other Resources (DC.Relation.Type)

Using the sub-element DC.Relation.Type, various relationships can be established to other resources. The use of this sub-element provides connections to other works in various relationships far beyond those used previously in the library catalog. For example: A photograph of an original Van Gogh painting

object is digitized. The metadata for the digitized image includes a pointer to a metadata set for both the photograph and (if it exists) to the original painting. This is important since the Creator, Date.Created and other information is unique. Such information was being embedded into one HTML-file for the digitized image, but it was found that it was more sensible to hold to the "one-to-one" principle: one set of metadata for each digitized or non-digitized object. This means that three sets of metadata would exist, with pointers in the respective DC.Relation field of each metadata set connecting the three. Especially the newly developed Resource Description Framework (RDF) (http://www.w3.org/RDF/) and the use of Dublin Core Elements in XML will facilitate bringing together the multifaceted aspects of information on digitized and non-digitized objects in a multidimensional information system.

The types of relationships to other resources which can be described using the DC.Relation.Type element are:

Inclusion Relations
- DC.Relation.IsPartOf
- DC.Relation.HasPart

Reference Relations
- DC.Relation.References
- DC.Relation.IsReferencedBy

Version Relations
- DC.Relation.IsVersionOf
- DC.Relation.HasVersionOf

Creative Relations
- DC.Relation.IsBasedOn
- DC.Relation.IsBasisFor

Mechanical Relations
- DC.Relation.IsFormatOf
- DC.Relation.HasFormat

Dependency Relations
- DC.Relation.Requires
- DC.Relation.IsRequiredBy

7. DC.Coverage

The DC.Coverage Element is the one Dublin Core element which requires sub-elements for differentiated use. One aspect of the DC.Coverage is the geographical aspect of being able to use geospatial and geographical coordinates to identify the location or physical coverage limitations of images (and also respective data). This aspect will be an especially important one for use in astronomy. For instance, a digitized image of Saturn might include the following metadata:

<META = DC.Subject, CONTENT = "Saturn">
<META = DC.Format, CONTENT = ".gif 640 x 512 pixel">
<META = DC.Type, CONTENT = "Image">
<META = DC.Date.Created, CONTENT = "19970623?">
<META = DC.Coverage.x CONTENT = " ">
<META = DC.Coverage.y CONTENT = " ">
<META = DC.Coverage.z CONTENT = " ">
<META = DC.Coverage.t CONTENT = " ">
<META = DC.Identifier, CONTENT = "http://www.not.iac.es/newwww/photos/images/satnot.gif">
where coordinates x, y, z give the specific location at the time of creation of the image, and t the duration of the exposure.

8. Who Produces Metadata?

The idea behind the Dublin Core and its emphasis on a simple structure was also to enable the creator of the document to submit his or her own metadata with the object itself (in the HMTL-header). Various templates have been developed for

this, including the Nordic Metadate Template (DC.Creator by Traugott Koch and Matthias Borell; http://www.lub.lu.se/dgi-bin.nmdc.pl), My Meta Maker for Theses (http://www.physik.uni-oldenburg.de/EPS/mmm/diss.html) which is also used not only for theses but with modifications for the German subject servers of the learned societies, the template for the German Educational Resources Server (http://dbs.schule.de/db/inconeue.html), etc. Ideally, in particular in the case of highly scientific material, the author is the best qualified to submit especially the content-oriented metadata for his or her work. Up until now, no adequate automated indexing service has been able to extract the necessary bibliographic information from HTML-files to fill in the respective metadata fields (even though some search engines purport to be able to do this and various projects to accomplish this have been proposed).

In addition to the author/creator, the publisher also can produce and provide metadata, especially in the case of articles, books, etc. (see above "DOI"). "Trusted Third Party Metadata" is a third alternative. Usually, the metadata from a trusted third party refers to more evaluative metadata (such as filtering mechanisms for children to avoid pornography, or evaluation of a certain teaching or learning program with regard to the school grade level, etc.), but may in the future also refer to "seals of approval" from certain learned societies or other recognized bodies for certain aspects of the necessary metadata.

Librarians and information scientists (as well as information specialists) will still be needed to provide metadata, even if certain metadata are provided by the author/creator and/or publisher. In some cases, it will mean augmenting the metadata provided by these other sources. In other cases, librarians may wish to sell (or provide in another modality) their metadata to publishers, or they may be involved in providing metadata for items for which no metadata are available (i.e., also for earlier works which need to be linked to newer ones – which is already part of retrospective digitization projects). This will also be an important task in order to achieve the ideal goal of metadata in terms of providing precise resource discovery while bringing together all items of related interest. Furthermore, other scientists and researchers can also produce or augment existing metadata for an object.

9. Metadata and Search Engines

Currently only a few search engines are capable of "harvesting" Dublin Core Metadata, though this number grows if the subject-oriented servers of the learned societies are considered. The Nordic Web Index (http://nwi.ub2.lu.se?lang=eng) is part of the Nordic Metadata Project and has been harvesting Dublin Core Metadata in HTML-documents for well over a year; the Australian Metadata Search Engine (http://www.dstc.edu.au/dgi-bin/RDU/Harvest/meta_simple_query.cgi) requires users to register their documents at the search engine headquarters so that their metadata can be "gathered" into the search engine index.

In Germany, the search engine "Fireball" (http://www.fireball.de) has its own set of metadata which it generates automatically, but this does not correspond to the Dublin Core set and is hardly more than the keyword and data metadata which Alta Vista and HotBot gather. Lycos and InfoSeek have not yet integrated metadata into their gathering criteria.

The more universal search engines have the philosophy that metadata are not yet being used widely enough to merit its integration in their search engines. Furthermore, the entire idea of for instance, Dublin Core Metadata, was to conquer the lack of precision of the general search engines and produce greater precision for resource discovery and retrieval. Hence, the search engines, which pride themselves on the high number of hits they can achieve, are logically not interested in reducing the number of hits to a fine 5-7 items (though the users might indeed truly desire this). However, this would limit their advertising facility greatly.

10. Conclusion

In summary, the advantages of metadata, specifically Dublin Core Metadata, include the fact that they can be integrated into the HTML-document header. They are easy to implement especially with the use of a template so that HTML-knowledge is not necessary; they enhance search precision and filtering; they allow bringing together resources which have formal, topical, or creative relationships to each other (→ multidimensionality); and they further standardization methods for information and use of Internet resources for scientific purposes.

The uses of metadata are vast – and perhaps just now being realized[1]. The implementation of Dublin Core Metadata has reached an internationally accepted standardization level. It has been officially accepted and is being used in the following countries:

Australia: The Australian Government has decided to use DC Metadata for all government information at all three government levels (local, provincial, national). The Distributed Systems Technology Centre belongs to the research and development core of the Dublin Core and has developed the Australian Metadata Search Engine, etc. The Australian National Archives are testing the Dublin Core for archival purposes (including description of physical objects, documents, digitized items, etc.). EdNA, the national educational server, uses Dublin Core.

Germany: Dublin Core has been accepted as the standard for the joint learned societies' servers (http://elfikom.physik.uni-oldenburg.de/IuK/) and has been recommended as the basis for all metadata developments in the German Digital Libraries Project (http://www.mathematik.uni-osnabrueck.de/projects/slot3/workshop98/akmeta2.html). The German Educational Resources Server (http://dbs.schule.de) uses Dublin Core both as HTML and as a database for its educational materials collection. The Southwest German Library Union is using the Dublin Core and a Dublin Core creating template for entering metadata on electronic publications in Germany. The Dissertations Online Project (http://www.educat.hu-berlin.de/diss_online/) of the German Research Foundation is implementing Dublin Core Metadata, as are other dissertations projects.

[1]In some implementations, Dublin Core Metadata have been used for: Preprint Servers, E-Journals (EurophysNet, MathNet, etc.), Dissertations (DDB, Dissertations Online (http://www.educat.hu-berlin.de/diss_online/), etc.), Texts, Library Catalogues (South-West Library Union, Germany, etc.), Teaching and Learning Materials (various educational servers world-wide, e.g., http://dbs.schule.de; http://gem.syr.edu), School Types and Directories (School Web (http://www.schulweb.de), University Departments and Course Offerings (MathNet, etc.), Multimedial Learning (German Digital Libraries Project), Arts and Museum Servers (Getty Museum, AHDS, etc.), Images (UCSB Alexandria Project, NASA, etc.), Archives (National Archives of Australia), Software (ELib)

Great Britain: All of the ELib-Projects (ROADS, SOSIG, OMNI and AHDS) have been or will be integrating Dublin Core Metadata in their services. The Arts and Humanities Data Services (http://ahds.ac.uk/) have belonged to the major developer group of the Dublin Core.

Nordic Countries: The Nordic Metadata Project has established the Dublin Core in Finland, Sweden, Norway, and Denmark with its own template (currently being translated into 8 languages) and search engine (http://linnea.helsinki.fi/meta/index.html). The Danish Government agreed with publishers to use the Dublin Core for all electronic publications as of January 1, 1998, and also as the basis for the Danish National Bibliography.

United States: Various libraries in the Digital Libraries Project are using Dublin Core Metadata. The ERIC Clearinghouse on Information Technology has created the Gateway for Educational Materials (GEM) (http://gem.syracuse.edu) using Dublin Core Metadata. Various museums are working on implementing Dublin Core Metadata for describing unique art objects in the networked environment.

Internationally, the Dublin Core is supported by the World Wide Web Consortium (http://www.w3.org/Metadata/) and the European Union endorses the Dublin Core in the metadata required by project proposals. The use of Dublin Core (not yet fully officially) in the DOI has already been mentioned.

Furthermore, the Dublin Core has submitted an RFC ("Request for Comment"; http://ds.internic.net/internet-drafts/draft-kunze-dc-02.txt) to the Internet Engineering Task Force for the simple 15 elements without qualifiers, on the basis of an HTML-embedment. A second RFC for the more structured, expanded Dublin Core (15 elements with qualifiers LANG, SCHEME, and subelements) is being currently prepared for submission. In addition, an RFC for the use of Dublin Core in XML is in the discussion and development stage and can be expected to be submitted within 1998. Furthermore, an attribute set for Dublin Core for Z 39.50 is being developed. A series of DC Working Groups are working together to precisely define the use and implementation strategies, as well as to give lists of standardized Type and Format categories, prepare a handbook, etc., see http://purl.org/metadata/dublin_core/. Several metadata registries to serve as reference servers for the Dublin Core are being developed at OCLC, in Germany (http://www.mpib-berlin.mpg.de/dok/metadata/gmr/gmr1e.htm) and in Thailand. Thus to conclude, metadata, especially Dublin Core, will no doubt increase in use, in particular in the scientific community, in support of improving search and retrieval conditions for information use of the Internet and could serve to be an important indexing tool for the electronic and non-electronic information sources in astrophysics as well.

11. Further Reading

An extensive reading list on Dublin Core Metadata can be found at
http://www.mpib-berlin.mpg.de/dok/metadata/gmr/dcmdlit.htm