National Library of Australia Staff Papers, 1997

Metadata

An overview

Paper given by Dr. Warwick Cathro, Assistant Director-General, Services to Libraries Division at the Standards Australia Seminar, "Matching Discovery and Recovery" August 1997

Introduction

What is metadata? My impression, from a number of recent meetings which I have attended, is that the concept is proving difficult to define with clarity. The Macquarie Dictionary defines the prefix "meta-" as meaning "among", "together with", "after" or "behind". That suggests the idea of a "fellow traveller": that metadata is not fully fledged data, but it is a kind of fellow-traveller with data, supporting it from the sidelines. My definition is that "an element of metadata describes an information resource, or helps provide access to an information resource". A collection of such metadata elements may describe one or many information resources.

It is inherent in the concept of metadata that there is an association of some kind between the metadata and the information resource which it describes. For example, a library catalogue record is a collection of metadata elements, linked to the book or other item in the library collection through the call number. Information stored in the "META" field of an HTML Web page is metadata, associated with the information resource by being embedded within it. The indexing data held by Web crawlers is also metadata (though not very good metadata) - linked to the information resource through the URL.

Metadata can be an information resource in its own right. For example, a review of a film - which on one level is a piece of metadata related to the film - is, on another level, a literary work with its own author and perhaps its own intellectual property constraints.

In recent years there has been a focus on metadata in relation to those information resources which can be accessed through the World Wide Web. I propose to concentrate in this paper on that aspect of metadata, and to discuss the Dublin Core metadata standard in particular. However, we should remember that there are other metadata schemes which are in use in relation to the Internet, and that metadata has a flourishing existence outside the Internet context. The huge amounts of cataloguing and indexing data created over many decades by

the library community, and the Abstracting and Indexing community, are equally entitled to be described as metadata.

The purpose of metadata

Whether in the traditional context or in the Internet context, the key purpose of metadata is to facilitate and improve the retrieval of information. At library school, we learnt to measure information retrieval in terms of recall and precision. If we miss a lot of relevant information, we have poor recall. If we get flooded by a lot of irrelevant information, we have poor precision. In certain circumstances (such as searches for patents) very high recall is essential. However, in most circumstances, searchers would be content with a small number of relevant documents, and would be willing to scan through a few dozen citations to identify them. Recall and precision factors of 10-20% are often acceptable for most purposes.

However, our own experiences with Web search engines frequently involve precision factors of much less than one percent. For example, a search of the World Wide Web using the search engine ANZWERS on the acronym "IETF" (which stands for Internet Engineering Task Force) retrieved 896,354 matches in early August 1997. Every Web page which mentioned the IETF in an incidental way was retrieved by this search.

This example illustrates that search engines can return a lot of irrelevant information because they have no means (or very few means) of distinguishing between important and incidental words in document texts. If we could target our searches onto words which are used as significant terms, we could achieve an enormous improvement in precision. Metadata can be used to achieve this by identifying just the major concepts of the information resource.

If we could target searches onto words or phrases that identify their correct role, we would also improve precision. For example, we could retrieve just those resources where "Green" is the name of the author, without retrieving resources about green peas or environmental issues. Metadata can be used to achieve this by identifying the different characteristics of the information resource: the author, subject, title, publisher and so on.

There is also a need to improve search recall - that is, to retrieve information resources that would otherwise be missed. For example, relevant information can be missed because sites contain types of resource in addition to HTML text, such as images, databases, and PDF documents. Metadata can support retrieval of these resources by identifying them, thus ensuring they are not missed by harvesting engines.

Recall can also be improved due to other factors. For example, it is known that most harvesting engines do not index every page on a site, but often only the top two or three hierarchical levels. Thus, these engines miss significant documents which, on larger and more complex sites, may be located in lower levels of the hierarchy. A better harvesting process would gather metadata from a repository created locally from a complete coverage of the local site. The data in this repository could then be gathered regularly by the harvesting engine.

Some search (or harvesting) engines do now take account, to some extent, of metadata stored within HTML documents, within the META field. The Search Engine Watch site, (http://searchenginewatch.com/features.htm) maintained by Calafia Consulting, documents the

behaviour of these search engines in this and other respects.

The Warwick Framework

There will always be a variety of metadata standards. Some standards have been developed to describe and provide access to a particular type of information resource, such as geospatial resources (an example of which is the FGDC standard developed by the US Federal Geographic Data Committee at http://www.fgdc.gov/Metadata/metahome.html). Other standards, such as Dublin Core, have been developed to provide a standard way of describing a wide range of different types of information resource, allowing these diverse types to be retrieved through a single searching process.

Before examining one particular metadata standard (the Dublin Core element set) it is useful to observe that an architecture has been developed to handle a variety of metadata sets. This architecture is known as the Warwick Framework, after a meeting held at the University of Warwick in early 1996. It has been described at http://www.dlib.org/dlib/july96/lagoze/07lagoze.html.

The Warwick Framework uses the "container-package" model. It provides a conceptual framework only: the method of handling the containers and packages must be provided by any particular application of the model. A container is simply any mechanism for aggregating packages. A package may be of three types:

- a "primitive" package, which contains one or more pieces of metadata; each primitive package has a type (for example, a MARC package, a Dublin Core package, an FGDC package)
- an "indirect" package, which refers to another information resource, for example through a link to its URL
- a "container" package: a package may itself be a container, and there is no limit to the degree of nesting involved here.

That is basically the Warwick Framework. It seems extraordinarily simple, but in fact it provides a powerful model for handling metadata. Its advantages are that it is modular (metadata is assembled in packages), extensible (there is no limit to the types of package which may be assembled in a container), distributed (through the use of indirect packages) and recursive (since a package may also be a container).

The Dublin Core standard

The Dublin Core metadata element set (http://purl.oclc.org/metadata/dublin_core) was developed during 1995 and 1996 as a response to the need to improve retrieval of information resources, especially on the World Wide Web. Dublin Core is being developed as a generic metadata standard for use by libraries, archives, government and other publishers of online information.

As originally devised, the Dublin Core standard was quite circumscribed in its aims. It was intended to be limited to describing "document like objects" such as HTML pages, PDF files and

graphic images. In practice it has proved to be difficult to define what this scope means, and particularly what it does not include.

The Dublin Core standard was intended to be descriptive, rather than evaluative. It does not provide for rating systems.

The Dublin Core standard was deliberately limited to a small set of elements which would have applicability over a wide range of types of information resource. There are currently fifteen elements in the Dublin Core standard.

Those who are attempting to implement the Dublin Core standard have raised a number of issues concerning both the semantics of the metadata (rules for the content of the fifteen fields) and the syntax (rules for structuring and expressing the fields themselves). In both areas, the Dublin Core standard is unstable, and is subject to reasonably rapid change.

The Dublin Core standard was intended to be silent on the question of syntax. To quote Stuart Weibel:

"syntax was deliberately left unspecified to avoid becoming bogged down in the tar pits of implementation minutiae "

(http://www.dlib.org /dlib/july96/07weibel.html).

In practice, of course, implementors of the Dublin Core standard have had to grapple with the syntax question. A syntax for expressing Dublin Core metadata in HTML pages has been developed (http://www.dlib.org/dlib /june97/metadata/06weibel.html) but it is subject to change as the HTML standard itself changes. In particular, the new standards embodied in HTML 4.0 should affect the metadata syntax.

The development of the Dublin Core standard is taking place through an informal international working group, whose activities are convened by OCLC in Dublin, Ohio. The group, which consists of a mixture of librarians and information technology professionals, has now held four workshops. The fifth workshop will be held in Helsinki, Finland in October 1997. The group also works through an active Internet e-mail discussion list. At this stage, Dublin Core does not have formal status as a standard either within the International Organization for Standardization (ISO) framework or the Internet Engineering Task Force (IETF) framework. However, a number of documents are being prepared as IETF Requests for Comment (RFCs). These documents will describe:

- the semantics of the fifteen element Dublin Core set;
- a method of encoding Dublin Core Metadata in HTML; and
- recommended semantics and syntax for the use of qualifiers with Dublin Core Metadata.

The Dublin Core elements

The following definitions of the fifteen Dublin Core elements are given at the Dublin Core Home Page.

TITLE	The name given to the resource by the CREATOR or PUBLISHER.
CREATOR	The person(s) or organization(s) primarily responsible for the intellectual content of the resource.
SUBJECT	The topic of the resource, or keywords or phrases that describe the subject or content of the resource.
DESCRIPTION	A textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions in the case of visual resources.
PUBLISHER	The entity responsible for making the resource available in its present form, such as a publisher, a university department, or a corporate entity.
CONTRIBUTORS	Person(s) or organization(s) in addition to those specified in the CREATOR element who have made significant intellectual contributions to the resource but whose contribution is secondary to the individuals or entities specified in the CREATOR element.

DATE	The date the resource was made available in its present form.
RESOURCE TYPE	The category of the resource.
FORMAT	The data representation of the resource.
RESOURCE IDENTIFIER	String or number used to uniquely identify the resource.
SOURCE	The work, either print or electronic, from which this resource is derived, if applicable.
LANGUAGE	Language(s) of the intellectual content of the resource.
RELATION	Relationship to other resources.
COVERAGE	The spatial locations and temporal durations characteristic of the resource.

RIGHTS	A link to a copyright notice or rights-management statement.

Semantics: the Minimalist/Structuralist issue

It was clear at the fourth Dublin Core workshop in Canberra in March that the development of this standard has reached something of a crossroads. As somebody commented at the time, this may be because the easy part (the definition of the core fields) has been completed. The key issue at the workshop was the extent to which the structure of the Dublin Core standard (until then, just a set of fifteen fields) should be elaborated. This is the so-called Minimalist/Structuralist debate.

For an explanation of the Minimalist/Structuralist issue, I commend the article by Stuart Weibel in the June 1997 issue of D-Lib magazine (The 4th Dublin Core metadata workshop report). To quote from this article:

"The Minimalist point of view reflects a strong commitment to the notion that DC's primary motivating characteristic is its simplicity. This simplicity is important both for creation of metadata (for example, by authors unschooled in the cataloging arts) and for the use of metadata by tools (for example, indexing harvesters, which will probably not make use of detailed qualifiers or encoding schemes). The goal of semantic interoperability across communities can only be achieved if there is a simple core of elements that are understood to mean the same thing in every case."

Weibel went on to contrast the contending viewpoint. "The Structuralists as a group accept the danger of [variability] in exchange for the greater flexibility of a formal means of extending or qualifying elements such that they can be made more useful for the needs of a particular community.... The underlying assumption of those who would deploy qualifiers is that the added structure and richness that they can provide can be used to good effect to enhance discovery. Every decision to deploy qualifiers should be measured against the question "Will this qualifier improve discovery?"

In this context, it is necessary to explain that there are three kinds of qualifiers. One kind, known as TYPE, refines the meaning of the field. Thus, "personal" and "corporate" are TYPEs which, if present, narrow the meaning of the CREATOR field. (This is usually expressed in so-called dot notation, as "Creator.Personal" or "Creator.Corporate"). Another kind of qualifier, known as SCHEME, explains the meaning of the data contained in the field. For example, "LCSH" is a SCHEME which helps to interpret the content of the SUBJECT field. These qualifiers - SCHEME and TYPE, along with a third one which denotes the language of the content of the field - are known collectively as the "Canberra Qualifiers".

A proposal for the use of qualifiers (both TYPEs and SCHEMEs) across the fifteen Dublin Core elements has been formulated by a group led by Rebecca Guenther of the Library of Congress.

This proposal may be found at http://www.loc.gov/marc/ dcqualif.html.

The following table summarises the position of this proposal. It should be noted that, for the most part, the SCHEMEs and TYPEs given in the table are regarded as optional. The word "no" means that no type (or scheme) is considered necessary for this field.

DUBLIN CORE QUALIFIERS PROPOSAL

FIELD	TYPE	SCHEME
TITLE	Main title	No
	Alternative title	
CREATOR	Personal	Yes (e.g. LCNA)
	Corporate	
	E-mail	
SUBJECT	No	Yes (e.g. LCSH)
DESCRIPTION	No	Yes (URL)

PUBLISHER	E-mail	No
CONTRIBUTORS	Personal	Yes (e.g. LCNA)
	Corporate	
	E-mail	
DATE	Created	Yes (e.g. ISO 8601)
	Last verified	
RESOURCE TYPE	No	Controlled list
FORMAT	No	Yes (MIME)
RESOURCE IDENTIFIER	No	Yes (e.g. URL, SICI)
SOURCE	Date	Yes (e.g. URL)

LANGUAGE	No	Yes (e.g. ISO 639)
RELATION	Parent	Yes (e.g. URL)
	Child	
	Member	
COVERAGE	Spatial	Controlled structure
	Temporal	
RIGHTS	No	Yes (e.g. URL)

The above table cannot yet be regarded as a consensus position of the Dublin Core community. For example, there is a minimalist view that virtually every one of the 15 Dublin Core elements should contain unqualified free text, without any TYPEs or SCHEMEs. Minimalists wish to see no formal constraints on the data content through adherence to particular standards or lists of authorised values, apart perhaps from the RESOURCE IDENTIFIER, which should contain a standard pointer such as a URL.

Structuralists would say that you can, if you wish, put free text or keywords into a field such as SUBJECT, but that you should also have the right to indicate that the keywords, phrases or codes in this field conform to a particular scheme, which might be a thesaurus (LCSH, MeSH, Art and Architecture Thesaurus, etc.) or might be a classification scheme (DDC, LC etc).

This example suggests a general approach to a compromise, based on the following principles. Structure should not be imposed on the minimalists. Equally, minimalism should not be forced on the structuralists. The key to compromise appears to be to agree that the default situation, at

least in most cases, is that the element requires no structure or qualifier. Thus, for SUBJECT, if you do not use any qualifier, then you are using free text keywords or phrases. But, if you wish to flag that you are using authorised subject terms or codes from a particular scheme, you are allowed to do so, and syntax will exist to support this. A further principle is that TYPEs should be used only where they clearly and significantly improve search precision. It is also essential that the definitions of all the elements are such that an unqualified value has a clear and useful meaning.

Specific structural proposals

In my view, probably ten of the fifteen Dublin Core elements could use unqualified free text as their default value, with a SCHEME being an optional addition. Something like five elements appear to require either a SCHEME or an authorised list of values as the default standard. One of these, RESOURCE IDENTIFIER, we have already discussed. The other four which probably require some structure in their default mode are DATE, RESOURCE TYPE, LANGUAGE and COVERAGE. The use of free text words in these five elements will probably fail to deliver satisfactory search precision. For example, a RESOURCE TYPE can be expressed in many different ways (article, paper, contribution, etc.) and without a controlled vocabulary the user will have to enter or guess all of the appropriate synonyms.

In June and July 1997 there was a flurry of discussion on the Dublin Core list concerning resource types. The result of this discussion has been a proposed list of authorised terms for resource type, developed by Roy Tennant at Berkeley. (This list can be found at http://sunsite.berkeley.edu/Metadata/types.html). The list is hierarchically structured, in three levels. These terms provide a controlled vocabulary which will improve precision by allowing users to confine a search to a particular resource type. While this list does not have universal acceptance, I believe that it provides a reasonable starting point for further development.

A structured approach has also been developed for the COVERAGE element, which describes the spatial and temporal characteristics of the information resource. This is another element for which improved search precision clearly depends on a clear and well defined data structure, and also on knowing what the data represents. The proposed standard (which may be found at http://alexandria.sdc.ucsb.edu/public-documents/metadata/dc_coverage.html) allows the spatial or temporal characteristics to be specified using either text or numbers (such as geographic coordinates and date-time).

With DATE we have the twin problems of defining which date we are talking about (date of creation, date of last modification, etc.) and of avoiding ambiguity in the field content due to different means of representing dates. It seems to be generally accepted that DATE would be best recorded using the ISO 8601 standard. However, there has also been discussion about extending ISO 8601 to allow for the inclusion of time, dates prior to the year zero, and dates after the year 9999.

The fourth Dublin Core workshop endorsed the idea of a Dublin Core Registry which would, amongst other things, support the registration of schemes that are cited in any of the Dublin Core elements. The schemes would be documented on the World Wide Web, allowing creators and users of Dublin Core metadata to retrieve information about the scheme, and to retrieve

lists of authorised values if necessary. A registration process implies that only certain organisations would be permitted to register particular schemes, or to delegate authority to maintain these schemes.

Conclusions

As the quantity of information on the World Wide Web multiplies rapidly, it will become increasingly difficult to retrieve information, with reasonable precision and recall, using the major search and harvesting engines. The use of metadata, combined with the use of improved harvesting processes, has the potential to improve retrieval of these information resources.

The Dublin Core standard has been developed as a universal general metadata set, which will support retrieval of a wide range of information resources. There is an active, current process to develop and refine the Dublin Core standard, and a number of projects which deploy the standard are now underway. There is a lively debate about the extent to which the standard needs to develop more rules about the content and structure of its data elements. My view is that a moderate amount of well-defined structure is necessary if we are to realise the underlying purpose of the standard - to enhance information retrieval.