Metadata: Enabling the Internet

Renato Iannella
Research Data Network CRC
DSTC Pty Ltd
Level 7, Gehrmann Laboratories
The University of Queensland, 4072, AUSTRALIA
Phone: 07 3365 4310 Fax: 07 3365 4311
Email: renato@dstc.edu.au

Andrew Waugh
Research Data Network CRC
CSIRO
Mathematical and Information Sciences
723 Swanston Street, Carlton, 3053, AUSTRALIA
Phone: 03 9282 2615 Fax: 03 9282 2600
Email: a.waugh@cmis.csiro.au

Abstract

Metadata is 'information about data'. That is, metadata describes some aspect of data on the Internet. There has been significant activity recently on defining the semantic and technical aspects of metadata for use on the Internet and WWW. A number of metadata sets have been proposed together with the technological framework to support the interchange of metadata. These initiatives will have a dramatic effect on how the Web is indexed and will improve the discovery of resources on the Internet by a significant factor.

This paper reviews some of the more popular Internet metadata systems and summarises the issues faced with moving towards supporting electronic metadata. The paper will also outline how the use of metadata will increase the level of precision and recall for WWW search engines, thus, &quotenabling the Internet".

Introduction

The information now available on the Internet on a particular topic varies greatly in both quantity and quality. The World-Wide Web (WWW) has enabled users to electronically publish information accessible to millions of people relatively easily, but the ability of those people to find relevant material has decreased dramatically as the quantity of information on the Internet grows.

One emerging trend is to enable the user to describe their own material with metadata. Metadata is &quotinformation about data". Metadata describes an Internet resource: what it is, what it is about, where it is, and so on. There are three major aspects for the deployment of metadata: description of resources, production of the metadata, and use of the metadata.

The first aspect which must be addressed is what set of information is to be captured by the metadata. This will depend on the type of the resource and on the purpose of the metadata. A metadata scheme must be sufficiently flexible to capture useful information about a wide variety of resources for a range of purposes. Ideally, a single metadata scheme should be used as this minimises the cost of using metadata. Unfortunately, it is unlikely that there will ever be agreement on a single metadata scheme and so a major aspect of metadata research is the relationship between different metadata schemes and the trade-off between the size and utility of the metadata element set.

The second aspect is the production of metadata. Metadata is essentially a summary of the data produced by various levels of &quotintelligence". Using humans to generate these summaries is expensive and metadata systems attempt to reduce this cost by making humans more productive by automating as much of the process as possible.

The final aspect of metadata concerns how the metadata is accessed and used. It must be retrieved in a form which can be processed with its semantics preserved. An important use of metadata is as a mechanism for resource location in distributed networks like the Internet. Metadata can provide information for the user to identify which resources they might be interested in. Once a resource has been identified, metadata provides the information to allow the resource to be accessed.

Metadata is not new. Librarians have been cataloging books and journals for hundreds of years. The library catalogue is, in effect, metadata that is used to find books and journals about a particular subject and to retrieve them from the library shelves. Although effective, library cataloging faces a scalability problem in producing the metadata. With so many dynamic documents being published on the WWW, it is not cost effective (or really possible) for librarians to professionally catalogue each WWW document.

What is Metadata

Metadata is information about information. Metadata has many uses in assisting the use of electronic and non-electronic resources on the Internet. These include:

- To summarise the meaning of the data (ie what is the data about).
- To allow users to search for the data.
- To allow users to determine if the data is what they want.
- To prevent some users (eg children) from accessing data.
- To retrieve and use a copy of the data (ie where do I go to get the data).
- To instruct how to interpret the data (eg format, encoding, encryption).
- To help decide which instance of the data should be retrieved (if multiple formats are provided).
- To give information that affects the use of data such as legal conditions on use, its size, or age).
- To give the history of data such as the original source of the data and any subsequent transformations.
- To give contact information about the data such as the owner.
- To indicate relationships with other resources (eg linkages to previous and subsequent versions, derived datasets, other datasets in a sequence, and other data or programs which should be used with the data).
- To control the management of the data (eg archival requirements, and destruction authority).

Metadata has an important role for supporting the use of electronic resources and services. However, many issues for effective support and deployment of metadata systems still need to be addressed.

Below is an example library catalogue record showing metadata about a book. It uses a (well-known) structure for the elements taken from The University of Queensland Library Catalogue (UQ, 1997).

```
Author:      Arnold, Eve.
Title:       Marilyn Monroe : An Appreciation
Publisher:   New York : Knopf, 1987.
Item Locn:   Central Quarto
Call No:     PN2287.M69 A74 1987
Status:      AVAILABLE
Descript:    141 p : ill. (some col.) ; 30 cm.
Subject:     Monroe, Marilyn, 1926-1962.
ISBN:        0394556720
```

Metadata is not limited to describing documents. Any resource (eg video, image, audio, etc) can be described with an appropriate metadata element set. The metadata below describes a satellite image of

Murray Bridge High School taken from the Environmental Resources Information Network (ERIN, 1997).

```
Title:     Satellite Imagery for Murray Bridge High School SA
Pixel Ref: x=244 y=243
Long:      139 17 08
Lat:       35 08 07
URL:       <http://www.erin.gov.au/ozglobe/satellite_images/
           gif_images/murray_bridge.html>
Last mod:  7 January 1997
```

Metadata Issues

The basic model used for metadata is known as &quotattribute type and value" model. Metadata is represented as a set of facts about the resource (eg &quottitle", &quotauthor"). Each fact is represented as an attribute (also known as an element). An attribute contains a type (which identifies what information the attribute contains) and one or more values (the metadata itself).

For example, the attribute "{Title} Marilyn Monroe : An Appreciation" has the attribute type &quottitle" (indicating that this is a title) and the value &quotMarilyn Monroe : An Appreciation".

Metadata standards, such as the Dublin Core described below, define sets of attributes which can be used to describe resources. These standards define:

- what information can be contained in the description (ie the set of attributes)
- which attributes are mandatory and which are optional
- what, precisely, each attribute means
- the syntax of the attribute value (ie rules for the format and construction of values). This might include sets of permitted values (ie a taxonomy).

However, there are a number of issues with this model, not all of which are resolved. These include:

- Accessibility of standards. Documenting attribute sets and building the information from the standards into systems so that the computer can assist the user in constructing and using metadata.
- How do different metadata standards relate? There are already many metadata standards and more will undoubtedly be created, which will lead to the situation where a resource will be described by two (or more) sets of metadata attributes. What happens if the two sets have contradictory information? Can metadata be transformed from one set to the other?
- How are metadata standards extended? It is often necessary to extend attributes sets to represent local information (eg linkages to existing systems). In addition, as new types of resources are defined, or new applications developed, it will become necessary to extend the metadata standards. How are these extensions published and incorporated into metadata applications?
- Internationalisation. The range of issues involved in extending metadata from the current English model extend from presenting values so that the preferred language is presented first, to the use of attributes which have no meaning in a particular culture.
- The linkage of data and metadata. Metadata needs to be tightly bound to the resource it describes. The metadata must be generated at the same time (or very soon after) the resource, modified when the resource changes or is deleted.
- Metadata is data (at another level of abstraction!). There are all the problems of storing it somewhere, finding it again, and understanding what the contents mean.

Metadata Specifications and Infrastructure

There are a number of emerging metadata specifications and deployment infrastructures that are gaining momentum on the Internet. This section summaries some of these metadata specifications which have a high WWW deployment. There are numerous other metadata sets available (eg IAFA Templates, GILS, etc) and deployment infrastructres (such as X.500 Directory Services). The challenge is to bring together the communities and agree to deploy flexible metadata infrastructures to support multiple (and extensible) metadata standards.

Dublin Core

The Dublin Core metadata set (DC, 1997) is intended to promote and develop the metadata elements required to facilitate the discovery of resources (documents and images) in a networked environment such as the Internet and support interoperability amongst heterogenous metadata systems. The Dublin Core working group participants cover a wide range people from industry, academic, research, and library communities. The current metadata set, which was finalised in December 1996, consists of 15 elements:

- Title
- Author or Creator
- Subject and Keywords
- Description
- Publisher
- Other Contributors
- Date
- Resoure Type
- Format
- Resource Identifier
- Source
- Language
- Relation
- Coverage
- Rights Management

Each element is repeatable and optional, and the entire set has been defined as *extensible*. A full description of the metadata elements can be found at <http://purl.org/metadata/dublin_core_elements>.

Each Dublin Core metadata element can also have a sub-type and sub-scheme information. For example, if an exisiting scheme is being used for *Subject and Keywords*, such as the Library of Congress Subject Headings (LCSH), then this information can also be attached to the element name. This provides additional semantics to the values of the metadata. If an exisiting type is being used, for example, a Uniform Resource Locator (URL), then this information is also stored with the element (in this case, the *Resource Identifier*). A complete list of proposed subschemes and subtypes can be found at (Knight & Hamilton, 1997).

This extra level of information then poses a problem of syntax and encoding for the Dublin Core metadata set. A number of syntax proposal have been made, including: RFC-822 Style Headers; SGML; MIME; and HTML META tags.

The latter option, HTML META tags, are fast becoming de facto standard (Weibel, 1996) as they are easy and quick to include at the beginning of WWW HTML files. An example encoding of Dublin Core using META tags is shown below. In this case, the metadata describes this paper.

```
<META NAME="DC.title" CONTENT="Metadata: Enabling the Internet">
<META NAME="DC.subject" CONTENT="(SCHEME=keyword) Metadata, Dublin Core,
                                  PICS, Resource Discovery">
<META NAME="DC.author" CONTENT="(TYPE=name) Renato Iannella">
<META NAME="DC.author" CONTENT="(TYPE=email) renato@dstc.edu.au">
<META NAME="DC.author" CONTENT="(TYPE=affiliation) DSTC Pty Ltd">
<META NAME="DC.author" CONTENT="(TYPE=name) Andrew Waugh">
<META NAME="DC.author" CONTENT="(TYPE=email) a.waugh@cmis.csiro.au">
<META NAME="DC.author" CONTENT="(TYPE=affiliation) CSIRO">
<META NAME="DC.publisher" CONTENT="(TYPE=name) DSTC Pty Ltd">
<META NAME="DC.date" CONTENT="(TYPE=creation) (SCHEME=ISO31) 1997-01-20">
<META NAME="DC.date" CONTENT="(TYPE=current) (SCHEME=ISO31) 1997-01-20">
<META NAME="DC.form" CONTENT="(SCHEME=imt) text/html">
<META NAME="DC.identifier" CONTENT="(TYPE=url) <http://www.dstc.edu.au/RDU/
                                    reports/CAUSE97/>">
<META NAME="DC.language" CONTENT="(SCHEME=iso639) en">
```

The Dublin Core working group also recognised that there would be more that just Dublin Core metadata being used on the Internet. An infrastructre was needed to support any metadata element set. This infrastructre is called the Warwick Framework and is a container architecture for aggregating logically, and perhaps physically, distinct packages of metadata (Lagoze *et al*, 1996). The architecture allows separate administration and access to metadata packages and proposes implementations of the Framework in HTML, MIME, SGML, and distributed objects.

Geographic Metadata

A more specific metadata set has been developed by the The Australia New Zealand Land Information Council (ANZLIC). The core metadata elements are for land and geographic directories in Australia and New Zealand. The ANZLIC metadata has particular categories (such as Dataset) with each category having a set of elements (in this case, Title, Custodian, and Jurisdiction). For more details see <http://www.auslig.gov.au/pipc/anzlic/metaelem.htm>.

Similarly, the Environmental Resources Information Network (ERIN) have developed a standard metadata set for their HTML documents (see <http://www.erin.gov.au/www-standards/metainfo.html> for more details). These include more specific metadata to describe spatial information, such as *Bounding Box* information, below.

```
<META name="North Bounding Coordinate" CONTENT="-9">
<META name="East Bounding Coordinate" CONTENT="154">
<META name="South Bounding Coordinate" CONTENT="-44">
<META name="West Bounding Coordinate" CONTENT="112">
```

PICS

The Platform for Internet Content Selection (PICS) developed by the WWW Consortium, is another metadata standard aimed at describing the content of Internet documents, in particular, the ratings on sensitive material (such as the level of nudity or violence). PICS does not actually specify any rating service, but the syntax and infrastructure to define such services (using labels). Third parties have already designed rating services (such as RSACi and SafeSurf). However, PICS also facilitates other uses for labels, including code signing, privacy, and intellectual property rights management. For more

information see <http://www.w3.org/pub/WWW/PICS/>.

Metadata Progress in the Internet

One of the major reasons for moving towards author-described resources with metadata is to try and provide more effective indexing services for the public. Current WWW indexing services will attempt to summarise a document by analysing the HTML code and producing a summary. This has not always produced effective results, and coupled with the enormous mass of WWW documents, the end result usually ends up disappointing the information seeker.

Some of the current WWW indexing services (in particular, AltaVista and Infoseek) are now currently supporting a limited metadata set (two elements; *Description* and *Keywords*). When these services index a WWW site, it will look for these META tags, and index the document based on the author-supplied list of *Keywords* (it will still try to index the rest of the document, but the keyword list takes precedence). The *Description* field is used to in the display of the returned result-set.

An example of the META tags for these two services is shown below.

```
<META  name="Description"
       content="The Resource Discovery Unit researches emerging
                technologies for the seamless discovery and retrieval
                of information and services on the Internet and WWW">
<META  name="keywords"
       content="resource discovery, Z39.50, X.500, urn, urc, metadata,
                information retrieval, WWW, internet">
```

When metadata becomes more common (either embedded in documents, such as the META tag in HTML files, or from a seperate metadata repository) and indexing services start to concentrate on indexing this information, we should see a marked increase in the effectiveness of information retrieval. The author-generated metadata (or even semi-automated) will add a higher level of quality.

The other advantage is that a user will then be able to do fielded searchs. For example, search for *Author=Smith* and *Subject=Metadata*. This would dramatically improve the search results, as it would ignore the documents where these two words appear just as free-text.

Conclusion

The use and promotion of metadata across a networked information environment poses significant challenges, although potential solutions are being developed. The potential solutions for *resource discovery* users will be to significantly improve the information retrieval problem that currently exists in the Internet. At the same time, provide tools to enable authors to better describe their resources.

Most work to date has been concerned with defining standard metadata attribute sets. Work is just starting on building systems which produce this metadata. Sufficient progress has been made, however, to make these standards potentially useful in finding information on the Internet.

The challenge of the next few years will be concerned with

- Integrating different metadata sets together
- Building metadata production and management tools
- Designing ways to extend metadata standards

Acknowledgements

The work reported in this paper has been funded in part by the Cooperative Research Centres Program, through the Department of the Prime Minister and Cabinet of Australia.

References

DC. *Dublin Core Metadata Set Home Page*, <http://purl.org/metadata/dublin_core>, 1997
ERIN. *Environmental Resources Information Network*, <http://www.erin.gov.au/>, 1997
Knight, Jon & Hamilton, Martin. *Dublin Core Qualifiers*, <http://www.roads.lut.ac.uk/Metadata/DC-SubElements.html>, 1997.
Lagoze, Carl & Lynch, Clifford A, & Daniel, Ron Jr. *The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata*. Cornell University Technical Report TR96-1593. <http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593>.
UQ. *The Univerity of Queensland Library Catalogue*, <http://library.uq.edu.au/screens/opacmenu.html>, 1997
Weibel, Stuart. *A Proposed Convention for Embedding Metadata in HTML*. Report from the W3C Distributed Indexing and Searching Workshop, May 28-29, 1996, <http://www.oclc.org:5046/~weibel/html-meta.html>