METADATA: A TOOL FOR CATALOGUING WEB RESOURCES

By

Pramod Kumar Singh*

ABSTRACT

As the Internet became an accepted source of electronic information, a variety of information communities have developed metadata to organize these resources to effectively serve their users. So metadata has taken on a more significant role of knowledge representation and data mining. In today's context, where the web contains the collection of massive heterogeneous objects, which need to be unified and linked in a single resource, we are witnessing both the growth of different metadata standards and the attempts to reconcile the common attributes in the existing overlapping standards. The goal is to access relevant information seamlessly, regardless of its type and location. This article addresses different aspects of metadata, the rationale behind it and gives a general picture of Dublin Core metadata elements.

---

* Student, Documentation Research and Training Centre (DRTC), Indian Statistical Institute (ISI), Bangalore-59. E-mail: pramod_15@hotmail.com

0. Introduction

As the World Wide Web continues to grow and expand, the amount of information redundancy or duplication within similar genres of sites will become increasingly evident and as Richmond said in 1999, there were over 100 million web pages available on Internet and are still growing exponentially. This fact serves to highlight the increasing difficulty of finding information on the Internet, like knowing the address or URL of a web site. Often this information is not readily available. At this point, search engines such as Google, Northern Light, Webcrawler and Rediff must be used. A novice Internet searcher may think, that search engine indexes all sites on the World Wide Web (WWW) and by matching of query words, they retrieve the relevant results. But unfortunately, this is not the case. Current search engines index only a fraction of available web sites and use their own set of algorithms to search through those sites.

If searching is, today, largely a matter of matching query words with the words in the text of articles, then anything that makes the matching process easier or more standardized is bound to improve the process. Metadata is one such tool, expected to improve matching by standardizing the structure and content of indexing or cataloguing information.

1. What is Metadata?

Metadata allow us to describe data sets, to advertise their existence to potential users, and to evaluate the fitness of data sets for use. They also enable us to find data in large collections and clearinghouses. That means Metadata describes an information resource in the haystack of a large collection of information.

The term "meta" came from a Greek word that denotes something of a higher or more fundamental nature. Metadata, then, is data about other data. Most commonly, it refers to descriptive information about World Wide Web and other networked electronic resources.

The concept of metadata is not a new one. Well before the first HTML page graced the web, millions of digital metadata records existed, created in semantic scheme known as the Anglo-American Cataloguing Rules (AACR), and stored in a framework called the MARC format. These records described the world's documented knowledge. These external metadata records, or surrogates, referenced resources existed separately from the resources they described. This traditional form of metadata deployment often allows the creator of the surrogate, the ability to customize the agent parsing the metadata. A common case is the cataloguer who not only creates the bibliographic record, but can also adjust the way the library's OPAC renders it. Metadata can be thought of in much the same way as the card catalogue, where each entry describes a resource in the collection. Where as, a library's card catalogue is located apart from the actual resource; metadata may also be included in the resource itself. It provides a user with a means to discover that resource exists, and how it might be obtained or accessed.

The term, metadata, is generally applied to electronic resources (though it doesn't have to be) and refers to "data" in the broadest sense. While the concept includes indexing and cataloguing information (information for "resource discovery"), it can go far beyond conventional document representations, such as MARC records.

In fact, because most search engines are text-based, it is essential to add a text description to non-text files if anyone is to find them. Databases of images are only beginning to be searched by non-text means, such as colour charts, or by matching faces or similar pictures. These non-textual databases abound in spatial information, not just geographic or political names, but coordinates of latitude and longitude, altitude, or depth; data that describe the forces of a tornado; infrared images of earth resources; images from Mars or the Moon; databases of famous musical themes, or museum collections.

In order to place metadata within its proper context, the following expanded definition can be used, which incorporates statements of functionality and environment: Metadata are the mechanism for both knowledge representation of digital collections and in data mining. It describes the attributes of a resource where the resource may consist of bibliographic objects (e.g. as represented by MARC metadata), archival inventories and registers (e.g. EAD metadata), geospatial objects (e.g. FGDC metadata), museum and visual resources (eg. CDWA, VRA, CIMI metadata), or software implementation (e.g. CORBA), to maintain just a few operational and proposed standards. While all these metadata formats differ in respective levels of specificity, structure, and maturity, their

primary purpose is similar: to describe, identify, and define a resource with regard to access patterns and filtering, terms and conditions for use, authentication and evaluation, preservation and interoperability.

2. Why Metadata?

In light of the unique and varied data contained in the databases, and the multiple purposes of these data, it is essential to construct an effective tool used to aid users in their searches. Traditionally, the retrieval tool for this purpose is a Controlled Vocabulary - an artificial language created for the uniform description, indexing, and retrieval of documents in a given collection. The primary purpose of the controlled vocabulary is to compel adherence to a standardized form of description of documents and of their subject contents.

The disparity among various databases and their varied uses negates conventional solutions for building an effective data retrieval system. A possible solution is to develop a metadata vocabulary that would support both access to the databases and to the individual records they contain. So the metadata vocabulary can be used as a common denominator for conceptually bridging across the various heterogeneous databases.

The major uses that can be thought for metadata, would be like: -

- to act as a surrogate for a larger database
- to characterize the original work sufficiently for the user to understand its contents, as well as its purpose and perhaps conditions of use
- to establish standard structure and terminology
- to provide information about an organization's data holding to data catalogues, clearinghouses, and brokerages.
- to provide information needed to process and interpret data to be received through a transfer from an external source, and
- to be a source for bibliographic data.

Metadata can help in ensuring the level of integrity of the data after necessary manipulations to preserve them for future use, e.g., conversion, transformation, migration,…etc.

There are some secondary benefits of metadata vocabulary, for users, like:-

- greater precision of results,
- fielded search,
- boolean support,
- less information overload.

Dispelling some common myths about Metadata :

- Metadata does not have to be digital. Library professionals have been creating metadata for as long as they have been managing collections. Increasingly, such metadata will be incorporated into digital information systems,
- Metadata relates to more than the description of an object. While many museums, archives, and library professionals are most familiar with the term in association with description or cataloguing, metadata can also indicate the context, management, processing, preservation, and use of the resources being described.

- Metadata can come from a variety of sources. It can be supplied by a human (a creator, information professional, or user), created automatically by a computer, or inferred through a relationship to another resource, such as a hyperlink.
- Metadata continues to accrue during the life of an information object or system. Metadata is created, modified, and sometimes even disposed of at many points during the life of a resource.

3. Search Engines and Metadata

Although initially both directories and search engines seemed to suffer from different types of problems, most of those difficulties were the result of ambitions that are likely to prove untenable in the long term. The web is simply getting too big for any single organization or service to catalogue or index, irrespective of whether they use people or computers to generate their indices.

Most of the search engines suffer from a number of serious problems, which affect both their ability to provide a comprehensive current index and the likelihood that users will find what they are looking for even if it has been indexed:

- The web crawling components are fully automated, which means that the web resources are selected by software rather than people, and are therefore variable in quality (i.e. Intellectual quality),
- Searching very large automatically indexed databases often results in extremely large result sets, which are frequently unusable despite increasingly sophisticated information retrieval tools, relevance ranking procedures, and context-aware artificial intelligence algorithms.
- Increasingly, information on the web is being generated "on the fly" from back-end databases (sometimes referred as "the hidden web"), which are beyond the indexing reach of the web crawlers.

To minimize these search problems, it is recommended to use metadata tags (i.e. metatags) to one's web contents. One goal of using metatags is to help the ranking of articles among the top sites in a list of retrieved "hits" when using an Internet search engine. The use of metatags improves the precision of our searches to the descriptor, identifier, author, title, or source fields; that means it can provide the field search.

Barring some innovative searching methods, the searching today is mostly a process of matching the query terms to the words in a document. If this matching is not perfect, then the relevant information will not be retrieved.

Proper use of indexing vocabularies and field structures, both in searching and in cataloguing, increases precision and minimizes the chances of false drops. Besides this, metadata also attacks three well known language problems that cause poor precision, that are:

- Polysemy, i.e. the problem of homonyms;
- Synonymy, i.e. the problem of many words representing the same concept, &

- ✍ Ambiguity - this problem can be solved by using the standardized metadata vocabularies.

Metadata not only improves precision, it can also help retrieval of pertinent documents by using the standardized term for each occurrences of a subject. Thus, a document will be retrieved from properly applied metadata even if it never uses the controlled term in its text.

4. Metadata Characteristics

There are three basic characteristics common to all metadata schemes: (i) Syntax, (ii) Semantics (i.e. Content), and (iii) Structure. A scheme's syntax can range from a highly complex format, such as the MARC record or the SGML encoded TEI header, to a basically unstructured scheme, such as the original DC. The semantics can include scores of complex data elements, whose contents is prescribed by the standards and rules, or it can have as little as two or three elements with no control at all over that content. Metadata can be contained in a variety of database structures (or architectures), including library catalogues, commercial database packages, or the recently formalized RDF standard. With increased importance placed on the global access to information and system interoperability, many of the new metadata schemes avoid the more complex syntax and rigid semantic content prescribed by the library cataloguing rules, MARC formats, controlled vocabularies, and traditional classification schemes. In some cases, however, metadata users and creators want more from their metadata than what a simple structure can offer. As a result, some schemes are being modified to include more elements, more data qualifiers and in turn result in a more complex structure. In addition, a few metadata sets (e.g. the DC and EAD) have added tagging conventions also, to support the use of authoritative data for names and subjects, and to indicate the authoritative sources for controlled headings.

5. Meatadata Schemes and Architecture

In the last five years, we have seen the rise of conflicting standards and projects for standardizing electronic resources. Some came from the library and research community, which has built new electronic standards on its original foundation. Others have emerged from groups that recognize the need for some sort of standard. While every metadata format is, in a sense, a "standard", many metadata formats have came up during last few years. The following list provides a sample of the better known metadata sets:

- ✍ Computer Interchange of Museum Information (CIMI);
- ✍ Federal Geographic Data Committee (FGDC);
- ✍ Dublin Core (DC) Metadata Element Set;
- ✍ EDUCOM Instructional Management Systems;
- ✍ Encoded Archival Description (EAD);
- ✍ Government Information Locator Service (GILS);
- ✍ IAFA Templates (IAFA/ WHOIS++);
- ✍ MARC Formats;
- ✍ Resource Description Framework (RDF);

- ✍ Text Encoding Initiative (TEI) Header;
- ✍ Visual Resource Association (VRA) Core Data, …etc.

Some of these metadata schemes are general in nature - such as the MARC Format or the DC - and are designed to accommodate information about electronic resources in a wide variety of disciplines. Other metadata schemes are more specialized, and apply to digital information in a specific format or within a specific discipline. What they all have in common is that they have a set of defined data elements that describe the entity and beyond that, they all vary as to the number of data elements, the content of the data elements, and the standards used, if any, for that content.

6. Dublin Core Metadata Standard

On the instigation of OCLC, Inc., the first workshop was held in Dublin, Ohio in 1995 to try to find a modus operandi in which the metadata can be formulated. In this workshop, they found consensus on a set of elements (now known as Dublin Core Elements), which is intended to be sufficiently rich to support useful fielded retrieval but simple enough not to require specialist expertise or extensive manual work.

The Dublin Core standard comprises fifteen data elements, which fall into three groups, which roughly indicate the class or scope of information stored in them: (i) elements related mainly to the *Content* of the resource; (ii) elements related mainly to the resource when viewed as *intellectual property;* and (iii) elements related mainly to the *instantiation* of the resource.

The Dublin Core Elements		
Scope	**Element**	**Description**
Content	Title	The name of the resource.
Content	Subject	The topic addressed by the resource.
Content	Description	A textual description of the content of the resource.
Content	Source	Objects, either print or electronic, from which this object is derived, if applicable.
Content	Language	Language of the intellectual content.
Content	Relation	Relationship to other resources.
Content	Coverage	The spatial location and/or temporal duration characteristics of the resource.
Intellectual	Creator	The person(s) or organization primarily responsible

property		for creating the intellectual content of the resource.
Intellectual property	Publisher	The agent or agency responsible for properly making the object available in its current form.
Intellectual property	Contributor	The person(s), such as editors, transcribers, and illustrators who have made other significant intellectual contributions to the work.
Intellectual property	Rights	A right management statement.
Instantiation	Date	The date associated with the creation or availability of the resource.
Instantiation	Type	The genre of the object, such as novel, poem, dictionary,…etc.
Instantiation	Format	The physical manifestation of the object, such as PostScript file or Windows executable file.
Instantiation	Identifier	String or number used to uniquely identify the object.

[*Source*: Metadata: An introduction, by Jan Smits, Cataloguing and Classification Quarterly, 27(3/4), pp.308-9.]

Although Dublin Core favours document - like objects (because traditional text resources are fairly well understood), it can apply to other resources as well. Its suitability for the use with particular non-document resources will depend to some extent on how closely their metadata resembles typical document metadata and also what purpose the metadata is intended to serve.
As its goals, the Dublin Core has the following characteristics:

- Simplicity of creation and maintenance;
- Commonly understood semantics;
- International scope; and
- Extensibility.

At the second DC workshop, held in Warwick, England, a conceptual foundation for an architecture for metadata was established, which is now known as "Warwick Framework". The Dublin Core-perhaps can be supplemented by additional metadata packages defined within the Warwick Framework- could be used to describe content where traditional cataloguing approaches are too costly, or where there is need to create metadata for contents that are not well served by current cataloguing practices.

7. Dublin Core in the Library World

Like most metadata standards, Dublin Core can be embedded in HTML documents to presumably enhance retrieval in search engines. The empirical effectiveness of META tags remains uncertain however. Search engine companies provide few specifics about the reliability of META-generated retrieval in Web pages, but most admit to indexing keyword META tags. Dublin Core is a rich structure, that will provide for very specific retrieval, if adopted by search engine proprietors. The motivation for AltaVista or Excite to adopt the Dublin Core syntax at present remains questionable. A finalized W3C metadata RDF should spur search engine companies into adopting the standard, and thus result in exact Dublin Core element targeting. How this will play out remains to be seen. Nevertheless, incorporating Dublin Core into library Web pages at present can only help retrieval. Additionally, individual library search engines can be crafted to target Dublin Core, thereby increasing retrieval for users of that specific site.

8. Conclusion

Whether you call it cataloguing, indexing, or metadata, the concept is a familiar one for Library and Information professionals, and to cope up with the exponential growth of this well-accepted source of electronic information, that is Internet, various information communities have developed metadata schemes to meet the needs of their users, and in turn each of the group, proposed a different standard. And even though each has its own merits, and is nicely applicable to the materials concerned to the group proposing it, we are left bare handed where no such internationally accepted standard is available which can be valid for all types of digitized materials.

9. References

1) http://purl.oclc.org/dc [Accessed on 7-10-2000]
2) http://www.nla.gov.au/nla/staffpaper/cathro3.html [Accessed on 7-10-2000]
3) http://www.dlib.org/dlib/january99/bearman/01bearman.html [Accessed on 27-11-2000]
4) http://www.ariadne.ac.uk/issue19/rowlatt/ [Accessed on 7-10-2000]
5) http://ukoln.ac.uk/metadata/desire/overview/rev_ti.htm [Accessed on 7-10-2000]
6) http://www.dlib.org/dlib/january00/01hodge.html [Accessed on 27-11-2000]
7) http://www.findarticles.com/cf_0/m1388/6_23/56979556/p1/article.jhtml [Accessed on 27-11-2000]
8) http://www.dlib.org/dlib/january00/chandler/01chandler.html [Accessed on 27-11-2000]
9) http://www.dlib.org/dlib/january99/buckland/01buckland.html [Accessed on 27-11-2000]
10) http://www.ukoln.ac.uk/metadata/review.html [Accessed on 27-11-2000]
11) http://www.ariadne.ac.uk/issue14/what-is [Accessed on 27-11-2000]
12) http://www. ukoln.ac.uk/metadata/publications/jdmetadata/ [Accessed on 27-11-2000]
13) http://www.findarticles.com/cf_0/m0FOX/1998_Oct_21/53141628/p1/article.jhtml [Accessed on 27-11-2000]

14) http://www.findarticles.com/cf_0/m0BLB/5_23/65803845/p1/article.jhtml [Accessed on 27-11-2000]