Creating metadata that work for digital libraries and Google

Alan Dawson, ***Senior Researcher/Programmer at the Centre for Digital Library Research, Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK***

Abstract

For many years metadata has been recognised as a significant component of the digital information environment. Substantial work has gone into creating complex metadata schemes for describing digital content. Yet increasingly Web search engines, and Google in particular, are the primary means of discovering and selecting digital resources, although they make little use of metadata. This article considers how digital libraries can gain more value from their metadata by adapting it for Google users, while still following well-established principles and standards for cataloguing and digital preservation.

This article introduces the concepts of functional and variable metadata, and explains why they may be of value to users and managers of digital libraries that rely on Web searching as a significant means of resource discovery.

Functional means something that works, so “functional metadata” is used here to mean metadata that fulfils its primary function of assisting information retrieval. Not all metadata does this in a Web-based world.

Variable means something that may change, so “variable metadata” is used to refer to metadata that may vary according to context. This is not the same as “dynamic metadata”, which has been used to describe educational metadata that can influence the behaviour of multimedia learning objects (El Saddik, 2000).

In order to consider why functionality and variability might be useful qualities for metadata, it is necessary to acknowledge the dominance of the Web and of Google as means of access and resource discovery for digital libraries. The current pre-eminence of Google extends well beyond the Web: a recent survey, drawing on users from 85 countries, rated Google as the world's number one brand name, above Apple, Mini, Coca-Cola, Samsung, Ikea and Nokia (Brandchannel.com, 2004). One might think information professionals would be delighted that an information retrieval tool had become the world's leading brand. However, some librarians have been known to denigrate Google because it “doesn't work”. Given that it can search millions of documents for thousands of users simultaneously, and deliver useful results within seconds, this is clearly a specialist interpretation of “doesn't work”. Yet one can understand the sentiment. Many of the things that librarians take for granted simply are not possible with Google. Trying to find articles written by Tony Blair, as opposed to those written about him, is difficult. A library catalogue system would make this easy, as “Blair, Tony” would be entered in the author field. But in a library system items are not normally catalogued at the article level, so the search might produce zero hits even though the retrieval system worked perfectly.

This basic problem illustrates the need for functional metadata (and the value of article-level retrieval). In the past cataloguers have been able to concentrate on capturing the metadata of an

object without necessarily having to consider how it might be of value to users. With digital libraries and the success of Google, all that has changed. Digital libraries need to make their metadata work as well as possible, which means adapting to a Google-dominated Web. Yet digital libraries may also need to follow well-established library standards for cataloguing and interoperability (perhaps as a condition of funding). The two goals may appear incompatible, but can be resolved by adopting a strategy to create metadata that is both functional (to optimise resource discovery) and variable (to optimise interoperability). This does not mean abandoning traditional cataloguing practices, but it may require new methods: Hyatt (2003) mentions "an emphasis on simplification" and "an increase in modularity and recombination of metadata" as significant current developments in cataloguing and metadata.

Functional metadata

In recent years a new profession of "search engine optimisation" has arisen, in which specialists advise companies how to help their Web sites rise through the rankings of Google search results. It is rather ambitious attempting to summarise the collective wisdom of this profession in three paragraphs, but it is worth mentioning the main points likely to affect digital libraries.

Firstly, the digital library domain needs to be indexed by Google. This is difficult to guarantee, but it does help to belong to a large well-established institution such as a university or public library. The domain name might not be cool or snappy, but that may be a price worth paying for high visibility.

Secondly, the digital library content needs to be search-engine friendly. This can be a problem for collections that are entirely database-driven, with pages created dynamically in response to user searches or selections. Using static URLs, and adhering to W3C accessibility standards, are good ways of helping make content accessible by search engines, as well as by users with disabilities.

Thirdly, the HTML <title> tag should have an accurate and specific entry for every item. This is important for two reasons: firstly, because the Google search algorithms give it significant weight, and secondly, because users see the contents of the <title> tag highlighted in their search results.

These three basic points seem a long way from the concerns of the metadata community, which for many years has been wrestling with the syntax of schemes such as qualified Dublin Core, IEEE LOM and IMS. Such issues are largely irrelevant when using Google. HTML does have a <meta> tag but its value is questionable; it has been misused by commercial Web sites and does little to enhance resource discovery. Sullivan (2002) concluded that the <meta name= "keywords"…> tag is effectively useless, as it is no longer used by most search engines, and that <meta name="description"…> is not used at all by Google, while Smith (2002), in a controlled study on the "Web impact factor" of electronic journals, found "little evidence that extent of metadata enhances the impact factor of the journal". This leaves <title> as the only effective metadata element in HTML, and the primary means of making metadata functional via the Web. So, how should digital libraries use the <title> tag? The answer is a little more complex than ensuring it contains an accurate title.

Variable metadata

Traditional library cataloguing is based on the concept of fixed metadata. A book may have a subtitle and alternate title as well as a main title, but these never change after publication. Electronic resources are more fluid, and so cataloguers sometimes add the date of viewing, but titles are still fixed in the catalogue record.

Why would anyone want a resource to have a variable title? The reason is to make its metadata more functional; different contexts require different metadata. This is hardly a radical concept. For example, most people in Scotland know that the highest mountain in Britain is Ben Nevis. The name is widely accepted and undisputed. Yet climbers may refer to it simply as "Nevis", while in Fort William (the town next to Ben Nevis), it is referred to as "The Ben", and some maps label it Beinn Nibheis. The established LCSH name is Ben Nevis (Scotland), to distinguish it from Ben Nevis in New Zealand. So, the name varies according to context and purpose. Redundant elements are removed as appropriate, and an international qualification is added in an international context. Librarians understand all this, which is why the concept of uniform title evolved. A MARC catalogue record might store several different forms of a title, but they are searchable collectively via a single title index. So the problem is solved as far as library catalogue records are concerned.

The problem for many digital libraries is that they are not accessed via a library catalogue; their content is usually discovered via Google, which emphasises the <title> tag. There is no <subtitle> or <alternatetitle> available, and no <author> or <subject> tag. Users have to browse numerous title tags (in their search results) to identify items of relevance. It therefore follows that the title tag is vital for digital libraries.

This situation is reminiscent of the library card catalogue before computerisation. Main entries were created in which the author and date of publication were appended to the title, so that users of the card catalogue could quickly see the most important metadata in one place. For non-book items, the physical form of an item (called the general material designation) was also included (e.g. sound recording).

This strategy can be adopted for the Google world in order to make digital library metadata more functional. There are no rules about the content of title tags in Web pages, so a title can look like this: <title>Communism and religion </title>; or could also include the item type, author and date of publication, like this: <title>Communism and religion [booklet cover] / Guy A. Aldred, 1911</title> which is far more useful amongst a long list of similar titles.

Some people may feel it breaks the usual metadata rules to put the author and date alongside the title. Yet the syntax of this example is closely based on the main entry point specified by AACR2 (the item type (booklet cover) gives finer detail than the standard AACR2 wording (electronic resource) that has little value in an environment where all resources are electronic). Although four fields appear concatenated in this example, this is merely a display format, derived from an underlying database in which the elements title, type, author and date are held in separate database fields, in accordance with basic data management principles. This is important, as Google may not be dominant forever, so alternative output formats might be required in future.

A further step in varying metadata would be to include the collection name as a prefix, e.g. <title>Red Clydeside: Communism and religion (booklet cover) / Guy A. Aldred, 1911</title>

This might make the entry even more functional when searching the Web via Google, but would be redundant when searching within the collection, where it would be unhelpful for every search result to begin with the collection name. The solution is to vary the metadata displayed, giving the collection name only when required. The feasibility of this depends on the method used for local searching. If it is based on harvesting then variability is more difficult, as the pages will have just one fixed title tag, but if based on dynamic database searching then output can be customised to produce whatever combination of collection name, title, author, date and so on is judged most useful to users. The difficulty is not technical but in deciding on the optimum amount of detail to return in the search results.

The potential value of variable metadata is not limited to titles. Subject terms are another obvious candidate for variability. For example, a standard taxonomy such as LCSH (required for interoperability), could be used alongside a controlled vocabulary of local variations (required for regional validity or subject specialists). Mapping between two taxonomies is not trivial (McCulloch, 2004), but does offer another means of maximising functionality of metadata.

The above examples illustrate the principles of variable metadata, but the practicalities of implementation are beyond the scope of this article. Suffice to say that generating different combinations of metadata elements for different contexts should be relatively straightforward provided the metadata is held in a consistent form in a structured and manageable database. Development of the Glasgow Digital Library (Dawson, 2004), which includes the Red Clydeside collection, has shown that putting theory into practice is certainly feasible, although further research and refinement of methods is continuing.

Conclusion

Library standards have evolved for good reason, based on sound principles, and should be respected and adhered to wherever possible. Long-term digital preservation requires items to be described accurately and consistently, using standard schemes and conventions. However, in a world where Google is the number one brand, a little more flexibility and creativity of interpretation might be a good idea for digital libraries. The use of functional and variable metadata can help users discover and quickly identify resources of interest from a long list of search results. It can also help digital library managers by increasing library usage while retaining quality of catalogue records and adhering to established international standards.

References

Brandchannel.com (2004), "Brand of the year survey results 2003", available at: http://brandchannel.com/start1.asp?id=195 (accessed 12 May).

Dawson, A. (2004), "Building a digital library in 80 days: the Glasgow experience", in Andrews, J., Law, D. (Eds),*Digital Libraries: Policy, Planning and Practice*, Aldershot, Ashgate Publishing.

El Saddik, A. (2000), "Metadata for smart multimedia learning objects", Proceedings of the 4th Australian Computing Education Conference, Melbourne.

Hyatt, S. (2003), "Developments in cataloguing and metadata", *International Yearbook of Library and Information Management 2003-2004*, Facet Publishing, London, available at: www.oclc.org/research/publications/archive/2003/hyatt.pdf (accessed 12 May 2004), .

McCulloch, E. (2004), "Multiple terminologies: an obstacle to information retrieval", *Library Review*, Vol. 53 No.6.

Smith, A.G. (2002), "Do metadata count? A Webometric investigation", Proceedings of the International Conference on Dublin Core for e-Communities 2002, Firenze University Press, pp.133-8.

Sullivan, D. (2002), "Search engine features for Webmasters", available at: http://searchenginewatch.com/webmasters/features.html (accessed 12 May 2004).