

Standards and quality deployment in digital libraries: the metadata role

Steve Ledwaba
National Research Foundation
P.O. Box 2600
PRETORIA, 001

Tel. (012) 481 4239
Fax: (012) 481 4239
E-mail: steve@nrf.ac.za

Abstract

Developing digital collections is both a complex and fascinating experience because of an array of opportunities it offers and the critical one being the need to maintain and sustain the collections metadata. The importance of maintaining metadata about physical materials cannot be overemphasised. The greatest challenge is the optimal management of the metadata because of the high expenses required. Many digital libraries and other archival services have been developed independently and this has led to the adoption and development of various metadata. Each system was designed and maintained by its own set of metadata. As a result, the metadata deployed were often not interoperable with other schemes. This became a barrier to the provision of and access to information across distributed databases. Metadata is regarded as the glue that binds the world of information together and efforts should be made to arrive at a standardised format.

The growth in electronic or digital resources has increased the need to search across different metadata structures simultaneously. In responding to this need, organisations and institutions were tempted to adopt the readily available metadata schemes. This was largely to minimise the expertise needed in developing such metadata and the maintenance thereof. An interface to other databases was then developed in order to complement their own collections. Furthermore, these institutions developed their own sets of metadata in order to manage and preserve their digital collections. This was done mainly with the use of proprietary software packages. In some cases mapping of descriptors from other metadata schemes deploying similar structures was preferred.

This paper attempts to discuss the development and maintenance of metadata for digital collections. Issues around the adoption of a common metadata format for digital collections will be interrogated and how the different types of metadata deployed in digital libraries add value to the notion of information and usage. Furthermore, questions around durability or sustainability of metadata and how organisations have attempted to customise metadata will be explored to forecast for the future optimisation of the metadata.

1. Introduction

The interdependence of libraries and database vendors in providing access to their information resources makes the adoption of standards a mandatory criterion that should underpin such activities and ventures. The growth in the production of information requires that the same information be available to users without access limitations. Advancement in Information and Communication Technology (ICT) has promoted electronic access over the traditional mode of information provision. Nowadays as more digital collections are developed, a standardised format to access these resources across distributed databases remains a challenge.

Metadata such as Machine Readable Catalogue (MARC) has been developed to represent physical materials. Various versions of this metadata such as SAMARC, USMARC and UKMARC were later developed in other countries. There is an abundance of metadata formats in existence today, some simple in their description while others are quite complex.

The growth in electronic or digital resources has increased the need to search across different metadata structures simultaneously without interoperability limitations. This article poses a critical question: How can a standardised metadata format that could be used to access information across distributed digital collections be developed? In formulating a response to this question, it is necessary to examine the following issues:

- What is metadata?
- Why a need for metadata?
- Which types of metadata are applicable to digital libraries?
- How is metadata managed in digital libraries?

Evidence gleaned from literature review gives us a better understanding of the concepts alluded to above. It also serves as a theoretical framework for addressing the standardisation challenge faced by digital libraries.

2. What is metadata?

According to Gilliland-Swetland (2000:1):

The term *metadata* is understood in different ways by the diverse professional communities that design, create, describe, preserve and use information systems and resources. As these communities come together to make the information age a reality, it is essential that we understand the critical roles that different types of metadata can play in the development of effective, authoritative, interoperable, scaleable, and preservable cultural heritage information and record-keeping systems.

Metadata is often referred to as data about data. Day (2001:11) postulates that the term 'metadata' is normally understood to mean structured data about resources that can be used to help support a wide range of operations. These include resource description and discovery, the management of information resources and long-term preservation (Day 2001:11).

Metadata is a structured description of an object or collection of objects. For the purpose of this paper, metadata is therefore defined as any data that can be used to identify, describe and locate an information product. *Metadata* describes how, when and by whom a particular set of data was collected, and how the data is formatted. Although different professions define *metadata* differently, the bottom-line is that metadata is expressed through information objects. Embedded in these information objects regardless of the physical or intellectual form, are the following features:

- *Content* relates to what the object is about
- *Context* indicates the who, what, why, where and how aspects associated with the object's creation
- *Structure* relates to the formal set of associations within or among individual information objects (Gilliland-Swetland 2000:1)

Information professionals are increasingly applying the term *metadata* to the value-added information that they create to arrange, describe, track and enhance access to information objects. Access to content has been the core of the library function. This is evidenced through the creation of bibliographic metadata such as indexes and abstracts. On the other hand, archival services have been focusing on context for preservation purposes. Examples of archival metadata include accession registers and catalogue records. For information to be organised and managed, it needs to be structured. Gilliland-Swetland (2000:3) argues that the role of structure has been growing as computer processing capabilities become increasingly sophisticated. These computer processing capabilities enhance searching and manipulation of information.

The creation and management of metadata is ever-growing. Libraries and abstracting and indexing services are continuously investing in the creation and management of metadata. On the same breath, publishers and other content providers are working on standards agreements to pave way for new forms of electronic commerce. It is worth mentioning that there is no single metadata that can be applicable in all the varying situations. With the competitive nature of the commercial sector and the proliferation and diversity of metadata, arriving at a standardised and interoperable form of metadata remains to be seen.

3. Why a need for metadata?

Metadata can come from various sources, that is, human beings, created automatically by a computer or inferred through hyperlinks. It attempts to answer questions such as:

- Who created this data?
- What time and location does this data apply to?
- What type of instrument and processing produced the data?
- What additional inputs were used to generate the data?
- What quality assurance has been performed on this data?
- Where the data is?
- What type of operations are possible on the data?
- Are there any access restrictions on the data?
- How individual data files are logically grouped into "collections"?

The key purpose of metadata is to facilitate and improve the retrieval of information. Metadata fulfils this purpose by verifying authenticity, process information in an appropriate format and order or pay online

The diversity of metadata as a useful tool has been widely acknowledged, for instance, the following functions have been identified by Gilliland-Swetland (2000:9):

- *Increased accessibility*: Metadata enhances information retrieval. Metadata can also make it possible to search across multiple collections or to create virtual collections from materials across distributed databases provided the descriptive metadata are the same.
- *Retention of context*: Libraries, museums and archival services maintain collections of objects that have complex interrelationships among each other and associations with people, places, movements and events. Metadata plays a significant role in documenting and maintaining these relationships.
- *Legal issues*: Metadata allows repositories to track layers of rights and reproduction information that exist for information objects. They also document other legal or donor requirements that have been imposed on objects.
- *Preservation*: For digital collections to survive, metadata that enables them to exist independent of the system that is currently used to store and retrieve them, need to be developed. Technical and descriptive metadata will be essential for the preservation purposes.

Metadata in digital libraries perform the following functions:

- *Resource discovery* – Metadata serves the same functions in resource discovery as in catalogues by:
 - Allowing resources to be found by relevant criteria
 - Identifying resources
 - Bringing similar resources together
 - Giving location information
- *Organising resources* – With the growth in the number of Web-based resources, portals are increasingly useful in organizing links to resources. Such links can be built as static web pages encoded with HTML. Hodge (2001:4) maintains that it is more efficient and increasingly common to build web pages dynamically from metadata stored in databases.
- *Interoperability* – Interoperability is the ability of multiple systems with different hardware and software platforms, data structures and interfaces to exchange data with minimal loss of content and functionality. There are two approaches to interoperability, namely, cross-system search and metadata harvesting. The z39.50 protocol is an example of the cross-system search. With z39.50 protocol, partners do not share metadata but map their own search capabilities to a common set of search attributes. Open Archives Initiative developed a protocol (OAI-PMH) which on the other hand requires partners to translate their native metadata to a common core set of elements and expose this for harvesting. The interoperability and exchange of metadata is facilitated by metadata crosswalks.

- *Digital identification* – Digital identification involves the inclusion of standard numbers, file name, URL or some more persistent URL (PURL) or the digital object identifier (DOI) in the metadata scheme.
- *Archiving and preservation* – Metadata is key to ensuring that resources will survive and continue to be accessible into the future. Hodge (2001:4) further argues that archiving and preservation require special elements to track the lineage of a digital object, to detail its physical characteristics and to document its behaviour in order to emulate it on future technologies.

As previously outlined, metadata is not only used for resource description and discovery purposes, but is also used to record any intellectual property rights, and to help manage user access. Metadata plays a very critical role in preservation and archival of digital resources.

4. Metadata for digital libraries

Digital libraries rely heavily on metadata for the organisation and management of digital resources. The existence of relevant metadata is the key to the future use of digital works. Digital libraries are collections of digital works that provide, among others, access to and preservation of resources. They are the “electronic extension” of functions typically performed to manage and access resources in a traditional library. Digital libraries help users to:

- Gain access to the holdings of libraries worldwide through online catalogues
- Locate both physical and digitised versions of scholarly articles and books
- Optimise searches through distributed databases (Sun Microsystems 2002:4).

The development of standards for digital libraries and networked resources is critical for establishing a common platform for access and preservation of these resources. Commercial interests have long dominated the issue of standardisation of access to information resources. There are three different types of metadata namely, descriptive, structural and administrative. These are essential for identifying, describing and locating information sources and reflect key aspects of metadata functionality.

4.1 Descriptive metadata

This metadata describes the intellectual content of a document or resource in a manner that facilitates a search by:

- Allowing the discovery of collections or objects through the use of search tools, and
- Providing sufficient context for understanding the results

Descriptive metadata is also known as “bibliographic data”. The description of an information product would include its title, what it is, who created it, when it was created, contributors, language, the subject, where it is located and so on. The main purpose of descriptive metadata is resource discovery and identification.

Machine Readable Catalogue (MARC) is an example of a descriptive metadata standard, and it has been used in traditional libraries for cataloguing books and other publications. However, its initial development was not meant for describing images, sound files and other multimedia types. Dublin Core metadata standard has been developed to cater for images and multimedia objects. Sun Microsystems (2002:16) affirms that Dublin Core was designed to provide a very widely accepted mechanism to allow discovery, with the option for different communities of users to adapt and customise it by adding more fields of particular importance to the community. New entrepreneurial metadata services which operate with browsers and search engines have emerged since the introduction of the Web.

4.2 Structural metadata

Structural metadata describes the relationships within or among related information objects. It defines the physical structure of a complex information object to facilitate information retrieval. An example would be the metadata that link a paragraph to a section and a section to a chapter within a book. In other words, the structural metadata would explain how individual pages make up a chapter and how chapters make up the book. In a digital library environment, an example would be to use metadata to link searches run on encoded text, embedded image, audio or video files to the pages where the hits occur. Structural metadata help users to navigate through individual objects that are part of a complex entity.

4.3 Administrative metadata

Administrative metadata are used in managing and administering information resources. They facilitate access, management and preservation of the digital resources. These metadata address issues such as who created the object, who has access rights, what is the file size, which viewer or player can be used, preservation and so on. Examples of administrative metadata include:

- Rights management relating to authorisation of access of copyrighted content available in digital form
- Technical metadata refer to information about the object's file characteristics or the capture or encoding processes used in creating the resource, that is, how a system functions or how the a metadata behave
- Preservation metadata relate to the preservation and management of information resources.

Metadata Encoding and Transmission Standard (METS) has emerged as an interoperability standard for managing and transmitting digital objects. The development of METS came about as a way of providing a framework for diverse digital collections available globally. However, the diversified nature of resource discovery, electronic commerce and rights management, call for a common understanding of data models if interoperability is to be achieved.

Figure 1 is an illustration of how different types of metadata are expressed in a digital library.



Figure1. The life cycle of objects contained in a digital information system (*Adapted from Gilliland-Swetland*)

The first stage of creation and multi-versioning involves entering or converting an information object into a digital format. Multiple versions of the same object may be created for preservation, research, dissemination or product development purposes (Gilliland-Swetland 2000:8). Upon creation of these information objects, they are automatically or manually organised into relational structures. These organised information objects would facilitate searching and information retrieval. This would be done by the system through creating metadata that track retrieval algorithms, user transactions and the system’s effectiveness in terms of storage and retrieval.

The user sifts through retrieved information objects to assess its relevance for the purpose of utilising, reproducing and modifying the object. Preservation and disposition relate to processes such as refreshing, migration and integrity checking to ensure continued availability of information objects. Inactive objects or information objects that are no longer necessary, could be disposed off.

5. Metadata management in a digital library environment

As discussed earlier, metadata serves as the pillar for the successful development and implementation of digital libraries. Companies and organisations have long been involved in the development of metadata for publishing their content online. This has been done for the purpose of commercial gain. As more companies became involved in content publishing, a wide range of software or databases which use proprietary metadata were growing at an alarming rate. Some of the proprietary solutions became “widely” used and as a result adopted as common standards for “most” of the databases. One of the reasons for using the proprietary solutions is that these solutions are readily available and there is no need for open standards which require resources and expertise.

The problem with proprietary solutions is that they tend to support specific databases and therefore fall short in promoting access to distributed or networked resources. As a result, there is a need to promote open standards whereby common set of metadata could be used. At the same time, a multiplicity of separate and functional metadata schemes should be developed to address complex metadata. The adoption of common metadata should avoid the dangers associated with reducing the standards to the common denominators which might exclude valuable datasets.

Dublin Core is a good example of the metadata which captures the essence of the information object. Its fifteen elements are applicable in almost all the information objects.

The escalating publication of content in an electronic format, coupled with the demand by scholars calls for access to these resources anytime and anywhere. The provision of such access has been the objective of the digital libraries. In essence, management and access to content can be expressed in the figure 2.

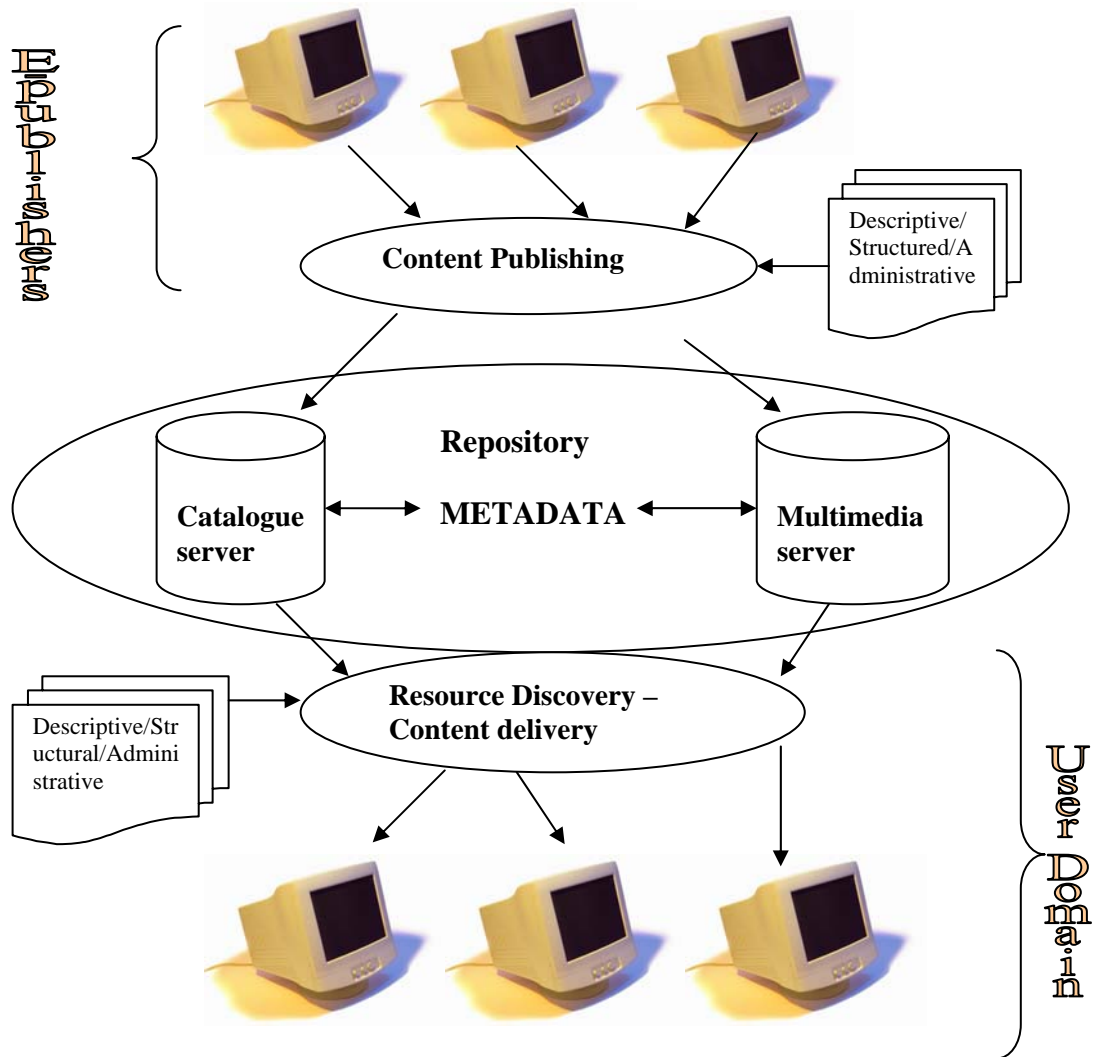


Figure 2. The role of metadata in a digital library

Apart from digitising content, an index that describes the content and facilitates information retrieval needs to be developed. The digital content and metadata are stored in a repository that has rights management capabilities to enforce intellectual property rights and e-commerce functionality needed to handle accounting and billing. Users will have access to the digital library through a browser.

6. Conclusion

Progress in the standardisation of metadata requires collaboration between metadata specialists and communities involved in rights management, resource discovery and archiving and preservation of information resources. The main challenge, however, is to get the cooperation of commercial publishers who have been developing proprietary solutions for libraries and other information services. An attempt to improve other metadata schemes such as MARC is a step in the right direction of striving for a standardised form of metadata, though this proved to be a slow and costly process. With more metadata available these days, it becomes necessary to verify their quality in terms of accuracy, integrity and trustworthiness.

The importance of metadata cannot be overemphasised. Of critical importance is the need to develop a set of metadata that can talk to a wide range of existing metadata and ensure archival and preservation of such resources. It is worth noting that different metadata schemes serve distinct needs and audiences. Complementary schemes can be used to describe the same resource for multiple purposes and for different groups. For example, a technical report could contain both bibliographic and statistical datasets. Representing this document for different groups of audience will therefore require the application of various metadata.

References

1. Allen, DY. 2001. Metadata primer for map librarians. Available WWW. <http://www.sunysb.edu/libmap/metadata.htm>. Accessed 20/09/04.
2. Cathro, W. 1997. Metadata: an overview. Available WWW. <http://www.nla.gov.au/nla/staffpaper/cathro3.html>. Accessed 11/08/04.
3. Coleman, A. 2002. Metadata: the theory behind the practice. 4th State GILS Conference, 26 April, Scottsdale, Arizona.
4. Day, M. 2001. Metadata in a nutshell. *Information Europe*, 6(2):11
5. Gilliland-Swetland, AJ. Introduction to metadata: setting the stage. Available WWW. <http://www.getty.edu/research/institute/standards/intrometadata>. Accessed 16/09/04.
6. EU-NSF Working Group on Metadata. Metadata for digital libraries: a research agenda, draft 10. Available WWW. <http://www.ercim.org/publication/ws-proceedings/EU-NSF/metadata.html>. Accessed 20/09/04.
7. Hodge, G. 2001. Metadata made simpler. Available WWW. <http://www.niso.org>. Accessed 09/07/04.
8. Payne, G. 1999. Standards, ever changing! Available WWW. <http://www.csu.edu.au/special/online99/proceedings99/104a.htm>. Accessed 09/07/04.
9. Sun Microsystems, Inc. 2002. Digital library technology trends. Available WWW. <http://www.sun.com/edu/libraries>. Accessed 09/07/2004.