

METADATA: The Foundations of Resource Description for Digital Libraries/Museums

Hsueh-hua Chen
Department of Library and Information Science
National Taiwan University, Taiwan

I. Preface

There are several organizations in Taiwan area which possess a lot of rare books, historical remains, artifacts and precious documents. However, the access of these collections seems unfortunately to be limited only to the minority of related researchers and scholars. This is against the overwhelming trend of information resources sharing. The most important thing is to distribute these valuable collections via the powerful WWW which will render these collections available to users all over the world in a more efficient way.

To achieve this aim, National Science Council of Taiwan (NSC) established a Taiwan Digital Museum Project (TDMP) since October 1998. Our project -- Resources Organization and Searching Specification, ROSS – is a sub-project under TDMP. ROSS aims to investigate metadata already developed; to meet the needs of the special characteristics of our rare collections, we will propose a proper model of information organization as a foundation for designing digital library system in Chinese environment.

This paper describes the definition of DL/M, the importance of metadata in DL/M, brief introduction to the development of Taiwan Digital Museum Project, as well as the primitive draft of our metadata (Metadata Interchange for Chinese Information, MICI) developed by ROSS team, and finally some problems need to be solved as well as the future development of metadata in this project.

II. Definitions of Digital Library/Museum

Some people think DL/M is equivalent to web pages or databases; but for the past two years, we discovered that DL/M exceeds far beyond that. DL/M has at least three purposes:

1. Culture preservation: Through digitalization, we may preserve precious cultural heritage and integrate historical remains, collections and researches from different institutions.
2. Academic research: Scholars may access and retrieve all the necessary information

through the web. In addition, we may promote researches on DL/M through related projects.

3. Education and learning: To provide educational media for the general public, for they are the majority of web users. The contents should be user-friendly and have educational values, so the general public will be able to utilize these materials.

DL/M is more than putting digitized documents or objects on the web for users to access/retrieve; it has the following two characteristics:

First, DL/M should be an extension of physical library/museum or an extension of its search system. It should be able to provide user-oriented functions of information storage, search, process and retrieve in an environment of multi-media with distributed access¹.

Second, the concept of a DL/M is not merely equivalent to a digitized collection with information management tools. It is rather an environment to bring together collections, services, and people in support of the full life cycle of creation, dissemination, use and preservation of data, information, and knowledge². Furthermore, this may expedite the cycle and accelerate the growth of knowledge.

In DL/M, in order for users to utilize information efficiently, information organization is a must. Traditionally, scholars in Library Science have been working on information organization for a long time, and they have well-established cataloguing rules, MARC (Machine Readable Catalogs), classification schemes, subject headings, indexes, abstracts, and so on. Thus, scholars and researchers in Library Science have considered themselves as specialists in this domain. Nevertheless, in the current web environment, many experts from computer science, and other disciplines are also interested in information organization. A lot of issues about DL/M were being discussed, and various types of metadata were developed to meet the needs of its specific domain.

III. Importance of Metadata in DL/M

Metadata is data which describe attributes of a resource, it is commonly used to refer to “data about data”³ or “data which describes other data”. It also could be defined as “additional information that is necessary for data to be useful”⁴, “information about the data that helps in optimization and management of that data”⁵, and so on.

Metadata describes a set of attributes of the collections. It is useful in several aspects of a

data system, including data access, data management and data analysis⁶. Traditionally, access function is carried out through card catalogs or MARC of Online Public Access Catalogs. Other tools such as indexes and abstracts are essential as well; and in a way, these are other formats of metadata. David Levy deemed that cataloging is a form of order-making; it is a set of practices which quite literally put a library's collections in order and provide access through a set of systematically organized surrogates.⁷ Cataloging provides attributes of the collections, such as title, author, and so on. Some attributes do not actually belong to the object being described, such as classification number, subject heading, etc. Nevertheless, they became the basis for library shelving, organization and information retrieval.

In DL/M, problems of hard disk storage and calculation capacity no longer exist. Under this circumstance, is it still necessary to use cataloging and surrogate (metadata) for information retrieval? Some people think since it is feasible to digitize the complete information, we should be able to put aside the surrogate and do the retrieval based on the information itself. Lynch (et. al.) objected this idea based on the following four points⁸:

1. Surrogates usually are much smaller than the primary objects, and they will be easier to process during information retrieval.
2. Concerning the intellectual property issues, information providers may be more willing to provide a surrogate, rather than the primary object.
3. The technology of content-based retrieval is still in developing and is not mature yet recently.
4. Surrogates may provide information in addition to the collection itself (i.e. subject headings, physical descriptions, etc.).

Metadata has the descriptive function of traditional cataloging. Its aim is to enable information managers and users to understand and identify information, and to further utilize and manage the information. In the environment of web information explosion, this need will be all the greater. From user's point of view, metadata supports the following five functions⁹:

1. location: to know the storage location of the resource;
2. discovery: to find out what resources are available;

3. documentation: to describe and to record features and contents of the documents;
4. evaluation: to assist users to determine value of the resource;
5. selection: to help users to decide whether to access this resource or not.

In addition, from the angle of system development, metadata has the following functions:

1. browsing and information retrieval

After describing features of the collections, one may organize these characteristics or attributes and arrange them in a particular order. There are two methods of information retrieval:

- a. browsing: System may be designed to arrange collections based on certain attributes, which in a way is familiar to users; thus enabling users to have an overview of the collections and to gain a sense of direction.
- b. retrieval: Metadata provides a basis for information retrieval. In the process of description, we will take out the essential information and organize them. In addition, we will define their semantics and establish their relations, which will increase the precision rate of information retrieval.

2. management

Management section of the metadata supports system management functions, i.e. system identification number.

3. to combine distributed objects and to present collections in a new way

After objects were digitized, they were stored in the form of distributed objects. An object might be an image or a full-text content, and so forth. For example, an archive may have two objects: one for the digitized image, the other for the full-text contents. Two objects are stored independently in the computer, and their handles are recorded in the metadata. When someone do a search on this archive, system will find its related objects, combining them, and present a complete collection to the user.

In the process of metadata development, information was analyzed and interpreted by librarians or content experts. Essential information were taken out or marked up, and this action gave strong semantics to the metadata, which our current automated technology is unable to do so. For example, system is capable to identify directly that the value in author

element represents a name rather than a meaningless string, which is very helpful in establishing relations among objects in the system. Therefore, we know that metadata plays a crucial role in DL/M. Disadvantage of full-text retrieval is lack of authority control; as a result, users may not be able to retrieve existing information or get some unwanted information. Thus, metadata and full-text retrieval should complement one another.

Various formats of metadata have been developed. Many DL/Ms developed their domain specific formats of metadata to describe attributes of their collections. For example, CIMI (Computer Interchange of Museum Information) , EAD (Encoding Archival Description) , GILS (Government Information Locator Service) , FGDC (Federal Geographic Data Committee Standard), TEI (Text Encoding Archival) Headers were developed for museum collections, archives, government publications, geographical-spatial information, and humanity resources, respectively. In addition, Dublin Core is regarded as the best potential candidate for digital information description and interchange format.

IV. Metadata Development of Taiwan Digital Museum Project (TDMP)

A. Taiwan Digital Museum Project (TDMP)

Recently, Internet is growing rapidly and its technology is updated continuously. There is a global trend to set up the web infrastructure, and information on the web is growing at an explosive rate. However, the popularity of web also caused some problems, and the most serious one is that the contents are too commercial oriented, which obviously lack of educational and cultural value. High quality contents will demand more advanced technology, which will be more difficult to establish; and content selection is very time consuming and cost is rather high. Nevertheless, our cultural tradition is long and heritage is abundant, and it is an excellent opportunity to promote culture heritage. While our government is promoting web infrastructure and web application, it should also think about how to establish our national style on the web.

Under the support of National Science Council of Taiwan (NSC), Taiwan Digital Museum Project (TDMP) was established last fall. In the global trend of WWW, TDMP is committed to establish cultural, artistic, scientific web sites to present relevant valuable resources through modern technology--not only to express our concerns/emotions, but to pass on knowledge/reasoning as well. Consequently, we may be able to disseminate quality contents of culture heritage and knowledge. Users may be free from the constraints of time and space; and through the use of WWW, users may enrich their life, expand their vision and enjoy a continuous learning experience. Besides, through the promotion of digital

collections, NSC hopes to stimulate the development of technology and industry for multi-media. TDMP has the following two main projects:

1. to build a working environment for cooperation

NSC invited related experts and scholars from Academia Sinica (AS), National Taiwan University (NTU), National Chi-Nan University (NCNU) and National Tsing-Hua University (NTHU), to establish a working environment for cooperation, thus to speed up the development of our national scientific and cultural educational web contents.

(i) Working Committee

It will be responsible for

- (1) managing whole TDMP
- (2) establishing a working environment for cooperation
- (3) promoting standards and guidelines for system development
- (4) proposing guidelines for content presentation and markup for digital resources

(ii) Education and Promotion

To train DL/M professionals through workshops and training courses. Furthermore, to present project results to the public through workshops and panel discussions.

2. to promote subject-based pilot projects

Through establishments of cultural artistic and scientific educational web sites to enrich the contents of WWW (which lack of cultural and educational materials). Scholars and experts are invited to write scenario based on the digitized rare collections of AS, NTU and National Natural Science Museum, and the aim is to produce web sites which can be utilized by the general public for both entertainment and educational purposes. In the first phase, our topics include: History of Dan-Shui River, Taiwan Indigenous People (Ping-pu tribal group), Taiwan Aboriginal Plants, Taiwan Fishery Resources, Taiwan Butterflies, Discovery of Chinese Characters, Culture of Han Dynasty, Firearms and Ming-Chin Battle.

B. Resources Organization and Searching Specification, ROSS

ROSS is a sub-project under TDMP. Its research scope covers the important issues concerning information organization and searching in DL/Ms in the Chinese environment, which includes information storage and management system design, user demand and information retrieval behaviors, and integration among different systems. Based on our research and past experiences, we think at least five topics must be addressed:

1. organizing digital information and establishing standards for Chinese resource description formats
2. analyzing user needs to develop a "user-friendly" environment
3. establishing thesaurus structure and authority file
4. designing systems for information retrieval and search service
5. integrating retrieval mechanisms of digital libraries/museums

Our current goal is to support every subject-based pilot projects, and our long-term goal is to formulate guidelines concerning resources organization and searching specification in the Chinese DL/Ms. These guidelines should be compatible with related international standards. Participation in international institutions (i.e. Consortium for the Computer Interchange of Museum Information, CIMI) may be helpful to the globalization of our DL/M systems in Taiwan. In addition, transparent integrated retrieval is an essential function, and it is a key issue in the globalization of Chinese DL/Ms. Therefore, it is very important to investigate and develop integrated retrieval system which meets the international standards; so that our system will be a system with interoperability rather than a closed system.

C. Establishing Metadata Interchange for Chinese Information (MICI)

Prior to TDMP, the Metadata Research Group was established under the project of National Taiwan University Digital Library/Museum (NTUDL/M) in March 1997 to study relevant metadata. Its responsibilities include understanding the features of collections, to study various metadata formats both domestically and internationally, to understand relations among metadata, database and the whole system, and to understand requests and retrieval behaviors of potential users. The Metadata Research Group held that, the format of metadata should be able to describe attributes of collections, to provide the mandatory access points to users, to have interoperability among different digital libraries so to be able to share information.

Most of the digitized collections of NTUDL/M were historical records, which included "Dan-Hsin File", "An-Li-Da-Chur Document", "Ino's Collection" and "Archives of the Dept. of Anthropology of NTU". After studying attributes of these historical records, the Metadata Research Group studied other related metadata, including CIMI, Dublin Core, EAD, TEI Header, and so on. However, due to culture differences and uniqueness of our collections,

these metadata cannot fully satisfy our needs.

In addition, with regards to interoperability, the Metadata Research Group has considered the possibility to adopt MARC (which was quite well established). However, after evaluation, we found MARC is too complicated for the historical records; not only that, MARC was mainly designed to describe books, and it cannot fully describe the attributes of our unique collections. For example, concept of "authorship" in historical records is not obvious; instead, "related person" is one of the crucial access points. If we put a particular information into similar (but not exact) elements reluctantly, it will result a loss in semantics, which is not desirable for us. Besides, in order to process MARC, we need to have software that is both specialized and complicated, which will become an undue burden for system design. Thus, based on the evaluation of cost and benefit, we decided to design a metadata suitable for rare Chinese collections; nevertheless, many features of MARC as well as other metadata were adopted.

In the process of designing metadata for historical records, members of the Metadata Research Group communicated continuously with content experts, end-users, specialists on user behaviors, and system designers. After much laboring, a draft of the metadata for NTUDL/M historical records was formulated in June 1998. After several months' testing, we started the revision in November 1998. The Metadata Research Group called several meetings to discuss how the metadata was used and how it should be revised. Finally, in the end of December, we reached a preliminary consensus.

The Metadata Research Group started to formulate metadata for other TDMP subject-based pilot projects' collections in addition to the historical records, which include historical objects, ancient maps, images and photos (for History of Dan-Shui River Project), and butterfly specimen (for Taiwan Butterflies Project). During the process of formulating our initial version, in addition to the discussions with content experts, we studied various metadata and web sites. In particular, *Handbook of Standards; Documenting African Collections* (published in 1998 by International Council of Museum, ICOM)¹⁰ provides guidance on the minimum amount of documentation required for museum objects from Africa, and we did a mapping of their elements to our historical object metadata. Concerning butterfly, a web site by US government was especially helpful (<http://www.npwrc.usgs.gov/resource/distr/lepid/bflyusa/wa/>). In the metadata format, elements were divided into seven areas: system management area, description area, subject area, resource type area, relation area, note area, reproduction area. We did a mapping for different types of metadata (see attachment 1), and we used Microsoft Access as the data entry interface (see attachment 2).

V. Problems and Future Development

Although the Metadata Research Group has some preliminary results; however, these are only domain specific formats for historical records, objects, ancient maps and pictures, and they are still subject to further development. There are some problems need to be solved. For example, we need to do a large scale users' information behavior study, in order to decide the appropriate access points. In addition, we need to discuss and decide which elements are mandatory or optional. As our projects go on, many tasks need to be maintained. Our future development will be:

1. to develop a user guide for MICI.
2. to develop and design metadata for other types of collections
3. to study attributes of other DL/Ms' collections, to update and improve our metadata continuously
4. to follow-up and study the development of existing mainstream metadata, such as CIMI, Dublin Core, EAD, and so on.
5. to discuss the feasibility for information exchange , for example, to do a mapping of our metadata to Dublin Core and TagG & TagM of Z39.50. This will allow the information exchange both domestically and internationally.
6. to study the feasibility of multi-lingual retrieval/access, so to be able to exchange information internationally.

In the future, we hope that our metadata will not only describe attributes of various pilot projects' collections under TDMP, it will also be consistent with user behaviors. We hope other similar institutions will be able to adopt our metadata working sheet to do their actual data entry; thus information will be interchangeable among different institutions.

References:

1. Maria Zemankova, "Workshop Report: Distributed Knowledge Work Environments (Digital Libraries, March 9-11, 1997, Santa Fe, New Mexico)".
DIGLIB@INFOSERV.NLC-BNC.CA

2. Maria Zemankova, "Workshop Report: Distributed Knowledge Work Environments (Digital Libraries, March 9-11, 1997, Santa Fe, New Mexico)". DIGLIB@INFOSERV.NLC-BNC.CA
3. Terry Kuny, "Metadata: What is it?" 24 Apr. 1997, DIGLIB@INFOSERV.NLC-BNC.CA
4. Terry Kuny, <Terry.Kuny@xist.com> "Second IEEE Metadata Conference," 2 Dec. 1996, <DIGLIB@INFOSERV.NLC-BNC.CA> (2 Dec. 1996)
5. A. Perkins, "Developing a Data Warehouse", <http://www.ies.aust.com/~ieinfor/dw.htm>
6. Marilyn Drewry, Helen Conover, Susan McCoy and Sara J. Graves, "Metadata: Quality vs. Quantity", Proceedings of Metadata Conference, '97.
7. David Levy, Cataloging in the Digital Order, Digital Libraries '95 <http://csdl.tamu.edu/DL95/papers/levy/levy.html>
8. Clifford Lynch, Avra Michelson, Cecilia Preston, and Craig A. Summerhill, CNI White Paper on Networked Information Discovery and Retrieval, Incomplete Draft <http://www.cni.org/projects/nidr/nidr.html>
9. Lorcan Dempsey and Rachel Heery, Specification for Resource Description methods. Part I. A Review of Metadata: A Survey of Current Resource Description Formats (March 1977) <http://www.ukoln.ac.uk/metadata/DESIRE/overview/>
10. ICOM – The International Council of Museums <http://www.cidoc.icom.org/>