Metadata Interchange for Chinese Information

Hsueh-hua Chen
Department of Library and Information Science
National Taiwan University
Taipei, 10764, Taiwan, ROC
sherry@ccms.ntu.edu.tw

Chao-chen Chen
Department of Adult and Continuing Education
National Taiwan Normal University
Taipei, 10764, Taiwan, ROC
cc4073@tpts1.seed.net.tw

Kuang-hua Chen
Department of Library and Information Science
National Taiwan University
Taipei, 10764, Taiwan, ROC
khchen@ccms.ntu.edu.tw

ABSTRACT

Accompanying with the growing Internet, DL/M has become an important researches issue. Metadata as a concrete foundation for Digital Libraries and Museums (DL/M) researches and systems, its role has being recognized by different research fields. Due to the essence of Internet, the paper not only describes the research and development of metadata for information, but also addresses on the issue of metadata interchange. We propose the metadata format MICI (Metadata Interchange for Chinese Information) developed by ROSS (Resources Organization and Searching Specification) team and design a metadata software tool, Metalogy, to fulfill all features touched on this paper.

I. Preface

Several institutions in Taiwan possess precious collections of rare books, historical remains, artifacts and precious documents. However, based on the consideration of preservation, access of these collections are very limited. Now, through the powerful WWW, we will be able to present these valuable resources on the WWW. In addition to the increasing public exposures, this will also preserve the life of the resource, which might be otherwise deteriorating.

Presently, mass amount of information can be obtained from the WWW. However, the quality of contents varies greatly. Thus, our first priority is to present these valuable resources on the web, make them available to users all over the world. Secondly, since web resources are not properly structured, our task is to organize these contents effectively based on different attributes of resources.

Many digital libraries/museums (DL/Ms) are developing metadata that is suitable for their digital information. To put digital information on the web, besides digitize the objects and contents, we have to establish cataloging system for information and develop organizing systems for resource, so to provide a more efficient retrieval mechanism. Obviously, traditional methods of information organization are not sufficient to deal with multi-media digital libraries, which contains various kinds of text, image, sound and interactive materials. Moreover, due to cultural and language differences, the metadata developed by other countries cannot be adapted to our collections directly.

Standardization is essential for the web, and we cannot create something that is only suitable for the metadata of one particular digital library. DL-resource integration is essential as well, and DL should provide a single interface with various information service options to the users. In addition, web resources cannot be provided by one library alone, and distributed retrieval system framework has been the mainstream for a long time. Thus, in organizing web information, we must take two factors into consideration: to integrate with current systems, and to be interoperable with other institutions for information exchange. Therefore, to develop an appropriate information organization model based on the attributes of our collections with reference to the in-depth study of other metadata, information organization systems is the first step to build the foundation of Chinese DL/M system.

II. The Development and Implementation of Chinese Metadata

National Science Council (NSC) of Taiwan established Taiwan Digital Museum Project (TDMP) in October 1998. Our project -- Resources Organization and Searching Specification, ROSS -- is a sub-project under TDMP [1]. The aim of ROSS covers information organization and retrieval issues about Chinese DL/Ms, which includes information storage and management system design, user-demand and information retrieval behaviors, and interoperability among different systems. Based on our researches 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

The current main task of ROSS is to support other pilot projects under TDMP, and its long-term goal will be to formulate guidelines for information organization and retrieval in Chinese DL/Ms. These standards should be compatible with international standards. Participation of international organizations (i.e. CIMI, Consortium for the Computer Interchange of Museum Information) may expedite the globalization of the Chinese DL/M systems. Likewise, transparent integrated retrieval is an essential function. Thus, to develop an integrated retrieval system that meets the international standards is crucial for us.

Prior to TDMP, ROSS Research Team (ROSS, in short) was established under National Taiwan University Digital Library/Museum, NTUDL/M) in March 1997 to study issues related to the Metadata Interchange for Chinese Information (MICI). Its responsibilities include understanding the history and features of collections; studying various metadata formats both domestically and internationally; understanding relations among metadata, database and the whole system; and understanding requests and retrieval behaviors of potential users. The ROSS held that, our metadata should be able to describe attributes of the collections, to provide the mandatory access points to users, to have interoperability among different digital libraries so to be able to exchange information, and to take consideration of the quality of description.

Most of the digitized collections of NTUDL/M are historical records, which includes “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, ROSS studied other related metadata, including CIMI (for art collections), EAD (for archives), and so on. However, due to culture differences and uniqueness of our collections, those metadata cannot fully meet our needs.

In addition, with regards to interoperability, ROSS 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 consideration of cost and benefit, we decided to develop a metadata for our

Chinese collections; nevertheless, many features of MARC as well as some other metadata were adopted.

In the process of developing metadata for historical records, members of ROSS 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 a five-month testing period, we started the revision in November 1998. ROSS 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.

Furthermore, ROSS started to formulate metadata for other TDMP subject-based pilot project collections, 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) 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 [2]. With regards to butterfly, a web site by US government was especially helpful [3]. In our metadata format, elements were divided into seven areas, and a mapping was done for different types of metadata. These sections are:

1. System Management Area: for the purpose of system management, which includes record number, cataloging language, language, cataloger, and cataloging date.
2. Description Area: for the purpose of describing the resource itself, which includes title, author, recipient, date of production, and place of production and use.
3. Subject Area: for the purpose of describing the subject of the resource, which includes subject/keyword, abstract, area of coverage, category by situation, category by function, category by material, category by technique, related event, related person, ethnic group, date, place, site, and cultural system.
4. Resource Type Area: for the purpose of describing the physical characteristics of the resource, which include type, physical description, and seal.
5. Relation Area: for the purpose of describing related resources, which includes collection, series, analysis, reference, and citation.
6. Holding Area: for the purpose of describing the acquisition and collection information of the resource, which includes owner, source, registration number, collection information, and rights restrictions.
7. Reproduction Area: for the purpose of describing the information format of the resource,

which includes digitized information and other media.

8. Note Area: includes general note, original description, condition, historical comments, reference, and publication record.

Under the coexistence of different types of metadata, in order to exchange information and to have interoperability among different systems, we developed Metalogy, a system which is able to manage various types of metadata based on the concept of Z39.50 of metadata management [4].

III. The Management, Maintenance and Exchange of Metadata

Due to different user-demand and collection attributes, DL/M's approach to information organization will be different as well. Consequently, different metadata was developed for different purposes. For example, Dublin Core metadata is designed for general web information [5]; FGDC (Federal Geographic Data Committee Standard) metadata is designed for geographic information [6]; CIMI (Consortium for Computer Interchange of Museum Information) metadata is designed for museum collections [7]; GILS (Government Information Locator Service) metadata is designed for government information [8]; and TEI (Text Encoding Archival) Headers is designed for archival materials [9]. Various types of metadata will facilitate institutions to organize resources properly; thus, users will be able to retrieve the needed information more effectively.

The development of various types of metadata shows that no single metadata can accommodate all types of collections and satisfy all kinds of user-demand. In the development of metadata for a particular kind of collection, one has to know the user groups and do a large scale of user group study. In addition, to meet users' needs and to describe the resource appropriately, a thorough understanding of the collection attributes is essential. In regards to the overall consideration of various kinds of metadata, one has to be knowledgeable about the interactions among different kinds of metadata, thus to establish an integrated DL/M system with the property of "distributed processing and integrated retrieval".

In order to achieve the above goal, it is crucial to manage and to maintain metadata effectively. According to the comprehensiveness of the content (from simple to more detailed descriptions), metadata can be seen as one layer of the hierarchical framework (please see Figure 1). The top-layer metadata contains the most basic "core elements", which is suitable for almost all resources and collections. The second-layer metadata is extended from the top-layer with some added elements. As a result, elements in the lower-layer metadata will be more specific, more detailed, and the user group will be much

smaller.

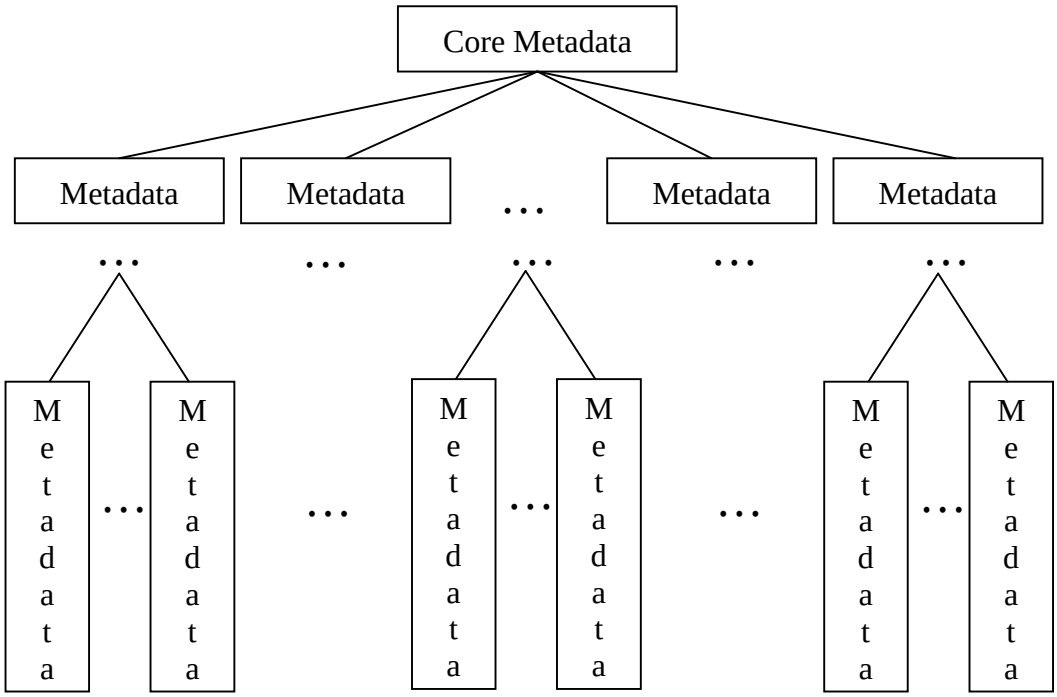

Figure 1. The hierarchical relationships among different types of metadata

For example, Dublin Core only has fifteen elements, which is the most common core metadata format. Other special domain (i.e. archival community) may add domain-specific elements below the Dublin Core. In this way, one will be able to exchange information with the most basic core elements with other field of study; and meanwhile, to exchange its complete elements (which is established with the cooperation of its own domain experts) within its domain (please see Figure 2). For example, EAD is more elaborated than the Dublin Core. If a system uses Dublin Core to describe resources, when it exchanges information with an EAD-metadata resource, EAD will be converted into Dublin Core through the process of mapping, and a degradation of metadata will result. However, this loss of information is inevitable.

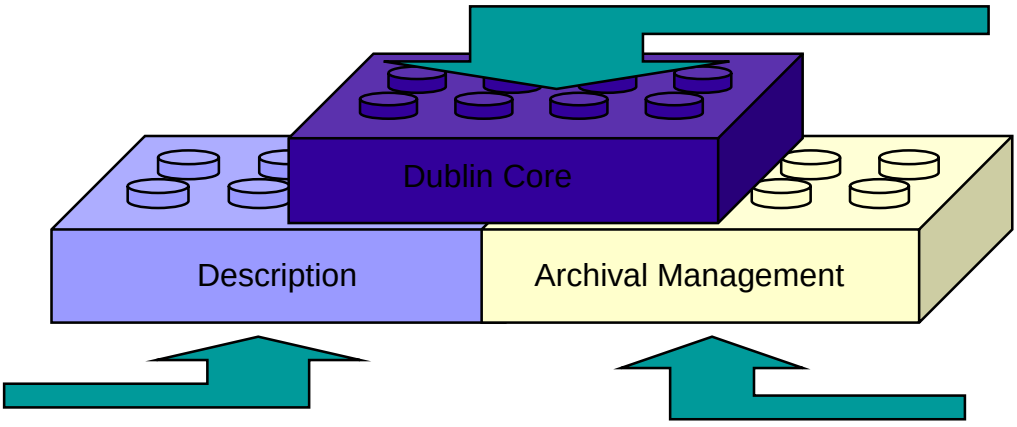

Figure 2. Concept structure of using Dublin Core as the core metadata

(Source: Eric Miller, John Perkins, Thomas Hoffman, "DC for Museum" (slide))

One institution may hold different kinds of resources. For example, libraries or museums may have collections on calligraphy/paintings, rare books, historical records, etc., and different metadata elements may be used to describe different attributes of the collections. For example, American Congress Library's digital library project proposed to use two formats of metadata to describe its resources: the traditional MARC and EAD. The metadata we developed for Chinese historical resources contains historical records, objects, pictures/photos, maps, etc. In order to manage different kinds of metadata conveniently and to utilize them with greater flexibility, Z39.50 was adopted. In other words, each type of metadata will be assigned a TagSet name. In Z39.50, two common TagSets were developed: TagSet-M and TagSet-G (where the Dublin Core is included in TagSet-G). Each institution of various domain can develop and name its own metadata: for example, the one developed by CIMI is named TagSet-CIMI, and the one developed by us is named TagSet-MICI. Nevertheless, MICI is a rather comprehensive metadata format, and not all resource types have to use all the elements. For better utilization of various kinds of metadata, the data-entry interface of MICI was designed in a way that users may select a metadata format from the following categories: historical records, objects, maps, and pictures/photos. We will talk about the details in the next section.

IV. Design and Implementation of a Multi-metadata Input-Output and Conversion System

This system consists of three components as shown in Figure 4. The first component is called Local Metadata; it serves to maintain all elements, institution-specific elements/database system schema, metadata input/editing, search/retrieval, data management, and the rights management of catalogers. The second component deals with the semantics and syntax of exchange of the standard metadata format. For example, conversion of XML-Dublin Core, XML-MICI, XML-MARC, or ISO 2709, etc. The third component has to do with Z39.50, which includes Z39.50 server and Z39.50 client. Its main purpose is to convert access points from Z39.50 client to database system-internal access points, and to convert the search results into Z39.50 record syntax, i.e. GRS-1.

A multi-metadata I/O and conversion system, Metalogy, was designed for the following two purposes: to utilize other metadata format, and to utilize MICI format with greater flexibility. The main concept behind this system is shown in Figure 3. Within the system, it has an all-element table, which is developed by authorized institutions. Core elements of the element table are the Dublin Core. The different institutions may select appropriate elements from this table, however, all core elements must be selected in order to ensure the

interoperability among different systems. One may use XML syntax to package the information, or one may consider using RDF to organize the information in the future.

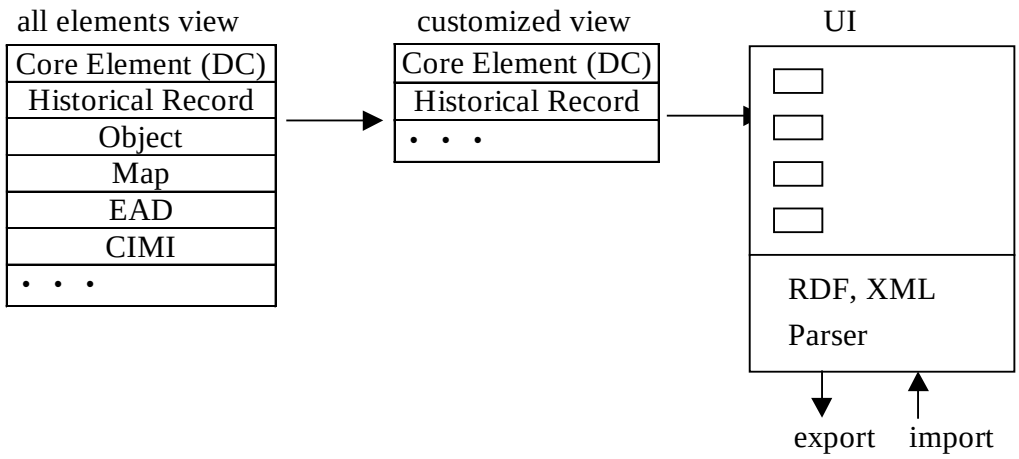

Figure 3. Framework concept of Metalogy

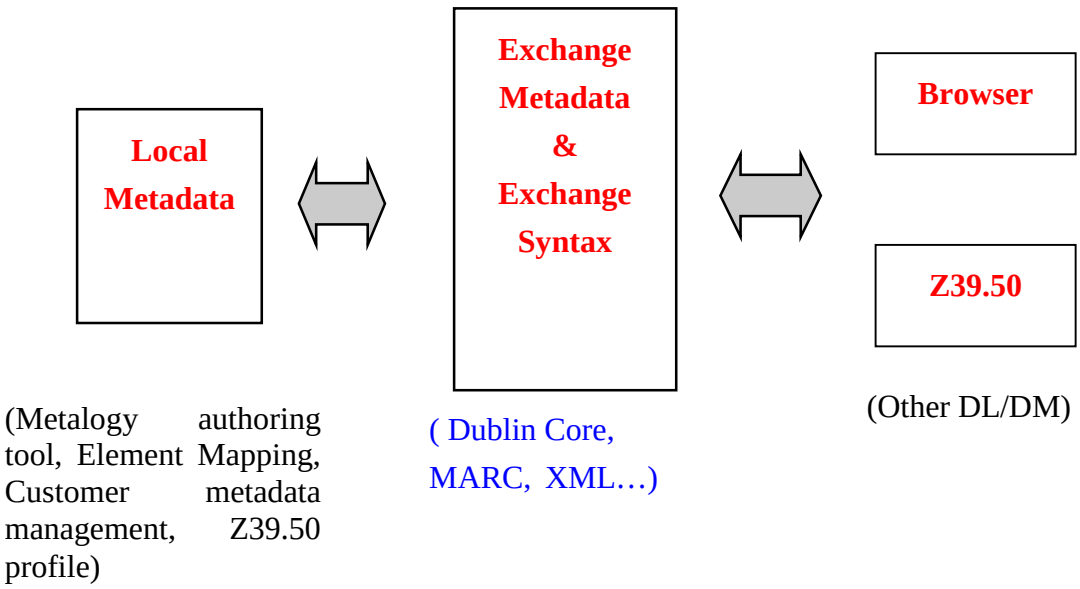

Figure 4. Three components of Metalogy

Currently, we have finished the first component of Metalogy. Its structure and modules are shown in Figure 5. The Dephi programming environment was used to develop this system. Figure 6 and Figure 7 are two snapshots of Metalogy. The catalogers could use the function shown in Figure 6 to customize the metadata format for special purposes. Then, they could key in the metadata using the function shown in Figure 7.

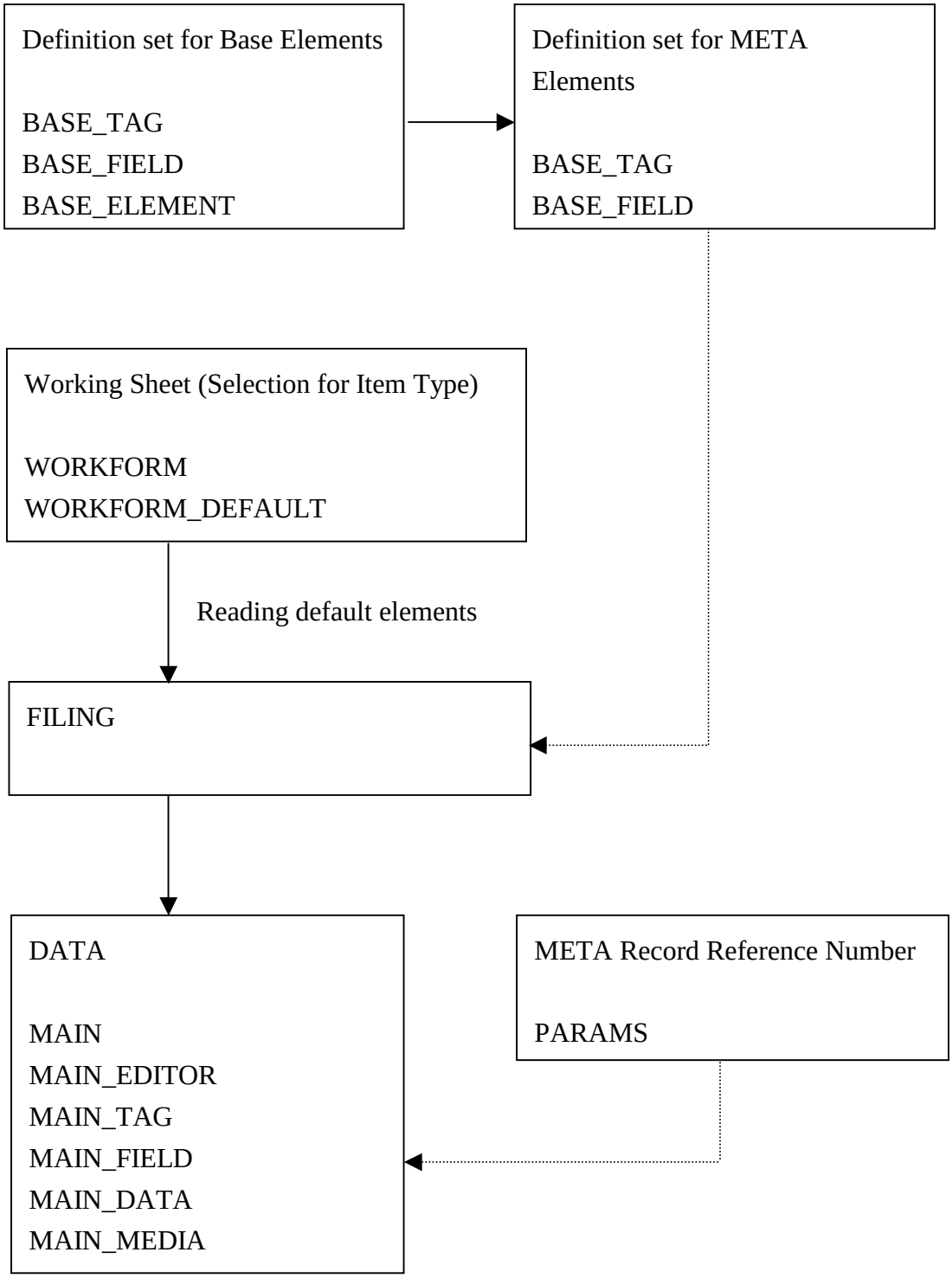

Figure 5. Structure and modules of the first component of Metalogy

Figure 6. Customization of metadata

Figure 7. Cataloging of metadata

V. Conclusions

Multi-metadata format has surely become an international trend for DL systems. However, from our experiences with other user institutions, we found that users prefer to have an interface with their schema that is relevant for their collections. In regards to user-demand, we should strive to simplify the database structure and the data-entry interface -- the only criteria is to be able to describe the collection attributes. However, in the long run, in order to exchange information, it seems reasonable to develop a multi-metadata format system with flexibility. Furthermore, implementation of Metalogy proves this is a feasible approach.

Acknowledgements

This work is supported in part by the National Science Council of R.O.C. under the contract NSC 88-2745-P-002-007. We would like to thank research assistants involved in this research.

References

[1] ROSS. “Resources Organization and Searching Specification – 1998.”
<URL: http://mail.lis.ntu.edu.tw > (in the present time)
<URL: http://ross.lis.ntu.edu.tw > (in the near future)

[2] ICOM – The International Council of Museums
<URL: http://www.cidoc.icom.org/>.

[3] USGS – Butterflies of North America
<URL: http://www.npwrc.usgs.gov/resource/distr/lepid/bflyusa/wa/>.

[4] H.H. Chen, C.C. Chen and K.H. Chen. “The Theory and Implementation for Metadata in Digital Library/Museum.” Journal of Library Science, National Taiwan University, No. 13, Dec. 1998, pp. 37-59. (in Chinese)

[5] Dublin Core Metadata Initiative, 1998, (13 Nov. 1998).
<URL: http://purl.oclc.org/dc/>.

[6] FGDC. “Content Standards for Digital Geospatial Metadata -- FGDC.” 1994,
<URL: http://fgdc.er.usgs.gov/> (13 Nov. 1998).

[7] CIMI. “Consortium for the Interchange of Museum Information – CIMI.” 1997,
<URL: http://www.cimi.org/> (13 Nov. 1998).

[8] GILS. “Guidelines for the Preparation of GILS Entries.” 1995.
<URL: http://gopher.nara.gov:70/0/managers/gils/guidance/gilsdoc.txt>.

[9] TEI Guidelines for Electronic Text Encoding and Interchange (P3)
<URL: http://etext.virginia.edu/TEI.html>.