World Library and Information Congress: 69th IFLA General Conference and Council

1-9 August 2003, Berlin

Code Number: 141-E
Meeting: 122. Library Theory and Research
Simultaneous Interpretation: -

A metadata lifecycle model for digital libraries: methodology and application for an evidence-based approach to library research

Ya-Ning Chen, Shu-Jiun Chen, Simon C. Lin
Computing Center of Academia Sinica, Taipei, Taiwan
[arthur, sophy, sclin@sinica.edu.tw]

Abstract:
Metadata is an emerging approach to organizing digital collections in order to enhance retrieval, preservation, and interoperability. The various issues confronting digital library projects that aspire to adopt metadata create the need for a systematic methodology. The Metadata Lifecycle Model (MLM), proposed in this paper, consists of ten steps: 1) acquiring basic metadata needs, 2) assessment of deep metadata needs, 3) review of standards and projects, 4) analysis of elements and standards, 5) preparation of metadata specification, 6) evaluation of metadata systems, 7) preparation of guidance and best practice, 8) development of metadata system, 9) maintenance of metadata service, and 10) evaluation of metadata performance. The MLM has already undergone two years of actual use, and a case study involving forty digital library projects has been conducted to examine its feasibility.

1.0 Introduction

Metadata is an emerging approach to organizing structured digital collections, in order to support precise retrieval, long-term preservation, and interoperability on an extraordinary Internet scale. Although there are many metadata practices in digital libraries, few literatures have addressed the methodology of a best practice for developing metadata. In the light of metadata provision, digital library projects often face a series of issues, including: how to get started, how to acquire metadata needs, how to choose a suitable metadata standard and adopt it, how to develop metadata specification, how to evaluate a metadata system, and so forth. A set of effective methods to develop metadata is thus very important. This paper introduces a Metadata Lifecycle Model (MLM) as a methodology of whole process of metadata provision for a digital library. In order to learn how well the MLM has been used, a case study of 40 digital library projects from the National Digital Archives Program in Taiwan (NDAP) examined the feasibility of this approach.

2.0 Literature Review

Metadata provision for a digital library involves various processes. Hudgins, Agnew, and Brown (1999) recommend a workflow for a metadata project based on project management perspective. This approach demonstrates ten tasks to manage a metadata project, including: understanding the entire project, documentation, maximizing existing infrastructure, choosing and evaluating the appropriate metadata standard, metadata record design, preliminary testing of workflow, initial staff design, workflow testing at midpoint, workflow testing at project conclusion, and reporting results. Facing the task of developing metadata for a digitized collection of the Kent State University Museum, Zeng (1999) focused on the issue of analysis and evaluation of the extent to which existing metadata formats can be applied to the historical fashion collection. The task involves 1) reviewing existing metadata formats and selecting one of these formats for the Museum collection; 2) augmenting this format as needed to meet the unique needs of the collection and its users; and 3) preparing metadata guidelines, examples, and a template for the museum registrar to develop and maintain the catalog for the entire digitized collection. Gilliland-Swetland's (1998) study on the relationship between metadata and the life cycle of objects contained in a digital information system shows that one can pay attention to metadata with a wider scope that includes the whole life span of digital collections, such as creation and multi-versioning, organization, searching and retrieval, utilization, and preservation and disposition.

Numerous attempts have been made by researchers to demonstrate how specific metadata formats are adopted and implemented at the digital library projects. The focus of metadata process includes:

- Methods of choice, evaluation, and adoption of existing metadata formats
- Methods of acquiring metadata needs in terms of lifecycle of digital objects
- Methods of preparation of metadata best practices
- Methods of development of metadata systems and tools

Let us now attempt to extend the observation into the idea of constructing a systematic way to present the whole metadata process for a digital library.

3.0 Theory: The Metadata Lifecycle Model (MLM)

3.1 The Concept of MLM

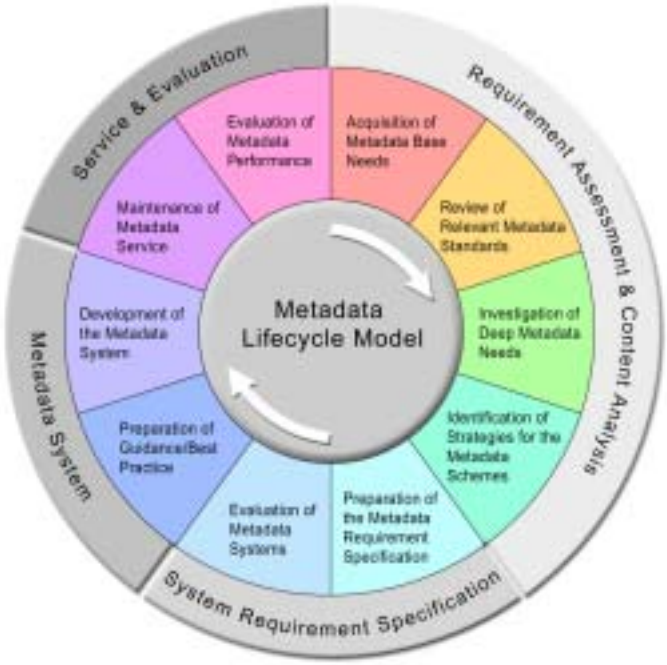
The MLM is a methodology involving a ten-step process by which digital library projects can design and implement metadata provision (See Fig.1). It can be reviewed again once the project requirements have been changed. The purpose of the model is:

- To achieve a consistent method of developing metadata for digital library projects, so that they might achieve greater effectiveness and better

quality.

- To conduct a content-based analysis for digital collections, which takes into account knowledge mapping and metadata research needs while adopting any existing metadata formats.

Fig.1 The Metadata Lifecycle Model



3.2 The Content of MLM

The ten-step of the MLM can be classified into four main groups as follows.

Group I Requirement Assessment & Content Analysis

Step1: Acquisition of Metadata Base Needs

Step2: Review of Relevant Metadata Standards and Projects

Step3: Investigation of Deep Metadata Needs

Step4: Identification of Strategies for the Metadata Schemes and Achieving Interoperability with Well-Known Metadata Standards

Group II System Requirement Specification

Step5: Preparation of the Metadata Requirement Specification

Step6: Evaluation of Metadata Systems

Group III Metadata System

Step7: Preparation of Best Practice Guidance

Step8: Development of the Metadata System

Group IV Service and Evaluation

Step9: Maintenance of Metadata Service

Step10: Evaluation of Metadata Performance

3.2.1 Acquisition of Metadata Base Needs

The first step of the metadata lifecycle is to interview the content experts or providers about their metadata requirements for each collection project, and analyze the attributes of collection projects through the four-layer for metadata selection (see Fig.2). The purpose of the interview is to acquire preliminary information and contacts, and also to establish better understanding between metadata staff and content providers regarding the project background information, such as purposes, goals and expected results. During the interview session, several points need to be clarified as follow:

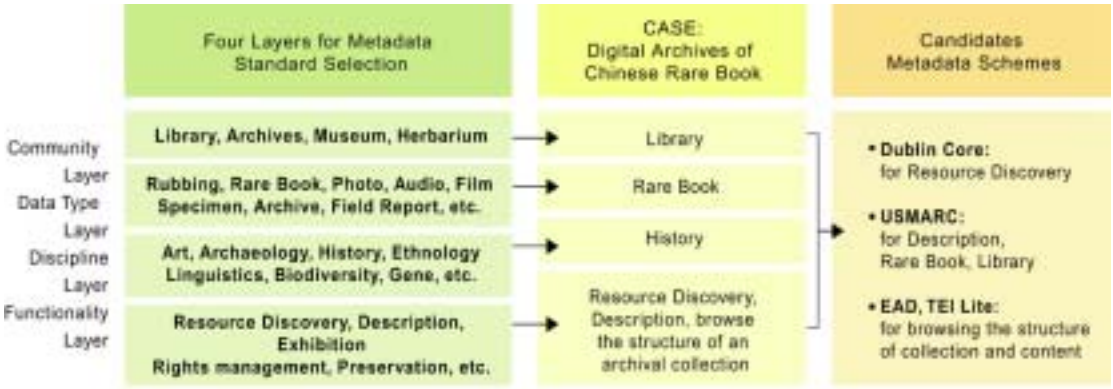
- *Contact information:* who is the contact window? Contact information of the project participants.
- *Metadata schedule:* when are metadata goals expected to be accomplished?
- *Metadata scope:* how many attributes of metadata are required for the projects? Such as attributes relating to thing, person, event, temporal, vocabulary control, and geographic name.
- *Legacy system:* basic information about the legacy system, including elements, structure, number of records, storage format, and input method and system. In addition, it is useful to understand the advantages and disadvantages of the system.
- *Metadata context:* Is only one metadata database constructed for this project? Are any other databases required to integrate with this metadata database, such as geographic information system (GIS)?
- *Metadata role and function:* What kind of metadata role is proposed for each project? What kind of function should be achieved by metadata, such as resources description, discovery, annotation, or administration?

3.2.2 Review of Relevant Metadata Standards and Projects

This stage involves identification of potential metadata standards, and examination of existing metadata schemes and their use cases. In the light of identifying potential metadata standards, the four-layer for metadata selection (see Fig 1) is developed to support the analysis of the collection project attributes, which can be a base to highlight the

focal point of metadata needs within the digital library.

Fig. 2 The Four Layers and Exemplar for Metadata Selection



After the four-layer for metadata selection is applied, the candidates of metadata schemes for a digital library are thus revealed. Current metadata trends and issues from around the world related to the profile of the digital library project can be collected and analyzed as a reference for practical application and future development. Through this stage the collection project members could well know what kind of differences exist with other similar or homogeneous collection projects, and also re-arrange the focus of expected project goals.

3.2.3 Investigation of Deep Metadata Needs

This step serves to identify a series of metadata needs of the digital library project more precisely. The concept of content analysis is deployed to acquire metadata needs from each use case through a set of work documents (WD). These include:

- WD1: Metadata Element – Name, Definition, Comment
- WD2:Metadata Element – Control Vocabulary
- WD3:Metadata Element –Example
- WD4:Metadata Element – Data type, Obligation, Maximum Occurrence
- WD5:Metadata Element – Unique Identifier System
- WD6:Granularity and Combination Diagram: Using the diagram shows the granularity of digital objects and their relationship. For example, fonts, series, files, and items in the archival collection.
- WD7:Association Diagram: An analytical context diagram with various relationships among metadata elements.

- WD8:Functional Requirements: eg. Input and display capacities of Chinese and Japanese Characters, default values, functional linkage
- WD9:Resource Discovery and Representation: including different levels of search (e.g.. keywords, advanced search), and different levels of display (e.g.. brief display, extended display)
- WD10:Metadata Administration

Through the set of work documents above, one may receive the benefits as follows:

- Metadata scope and context are clarified, and related relationships are clearly drawn and attributed to a diversity of categories.
- It could ensure what kind of systems and databases are integrated by metadata mechanisms such as GIS.
- Examples for collection projects are collected as a basis for the best practice.

3.2.4 Identification of Strategies for the Metadata Schemes and Achieving Interoperability with Well-known Metadata Standards

This stage involves the formulation of metadata strategy for the digital library, based on the previous findings. The strategy usually includes adopting one or several existing metadata standards, or developing a suitable metadata scheme based on some standards. It is common knowledge that there is no single accepted standard for cataloguing all collections in digital form. This may explain the reason why metadata application profiles are more and more popular. One example is the metadata application profile based on IEEE LOM (Institute of Electrical and Electronics Engineers, Learning Object Metadata). In addition, the mapping work should be done between the developed metadata scheme and international metadata standards, such as Dublin Core and The MAchine-Readable Cataloging format 21 (MARC21).

3.2.5 Preparation of the Metadata Requirement Specification

To achieve a common agreement among collection project participants, metadata specialists, and system designers, a Metadata Requirement Specification (MRS) should be prepared as a bridge for these members from a variety of disciplinary domains. The specification contains several components: executive summary,

background information of the collection project, project participants, objectives and scope of the metadata system, statement of adopted metadata standards, metadata elements and structure, attributes of metadata elements (such as label, length, data type, index key, and so forth), an input template, related standards mapping (e.g.. mapping Dublin Core), a context diagram for metadata scope for a range of systems, functional requirements (e.g.. Input and display capacities of Chinese and Japanese Characters), control vocabulary lists, XML DTD (Extensible Markup Language, Document Type Definition), and so forth. Further, the specification is also considered as a ground for different purposes:

- A confirmation of project requirements for metadata by related project participants.
- A communication bridge between metadata team and system designers.
- A basis by which collection project members can verify metadata system and function.
- A groundwork for a best practice and a crosswalk mechanism for existing metadata standards.

3.2.6 Evaluation of Metadata Systems

This stage involves the evaluation of potential metadata systems. The collection project members can select an existing system developed by homogeneous or similar collection projects. Alternatively, the collection project members can decide to develop by themselves or in collaboration with others, such as university or industry teams.

3.2.7 Preparation of Guidance/Best Practice

This stage involves generating best practice guideline for individual metadata elements that are given within the MRS. It can be used as a checklist and reference when collection project participants apply the metadata scheme. It can also ensure the quality control of the metadata records in the collection project. The content of guideline includes the definition, description rules, system suggestion, examples, and mapping to relevant standards for each metadata element.

3.2.8 Development of the Metadata System

The metadata development task is transferred to system developers at this stage. They develop metadata tools and systems according to the MRS. During the time of system implementation, staff from among collection project teams, metadata teams, and system development teams may need to communicate and discuss continually to ensure an effective way of system design. Both collection project and metadata teams will carry out the evaluation of metadata system and feedback for modification, once the system has been finished.

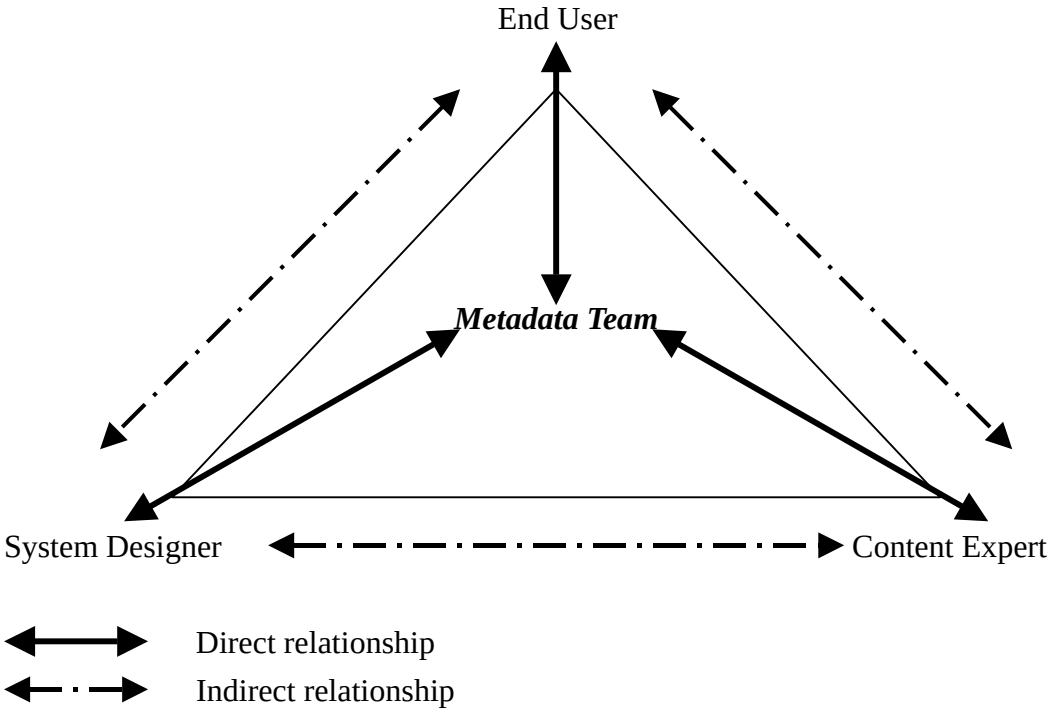
3.2.9 Maintenance of Metadata Service

The purpose of developing metadata services is to guarantee the quality assurance of metadata mechanisms. A service model is constructed to formulate the service items. The metadata service model is composed of three basic elements as follows (refer to Figure 3)

- Service Mechanism
  - Enhancing user interface and related function for collection projects or end users.
  - Providing consultation in delivering knowledge of metadata and content analysis to content providers.
  - Revising the functional requirements of metadata systems. For example, one collection project team may wish to modify several functions in the metadata system after testing the system.
  - Offering advice on issues of interoperability across a wide range of metadata standards, e.g., crosswalk, meta-search, and so on.
  - Providing and harvesting metadata records to achieve metadata interoperability
  - Accomplishing a crosswalk mapping among the legacy records, project required metadata fields, and selected metadata standards for system designers.
  - Providing own language version of metadata standards such as Chinese version of Encoded Archival Description (EAD)
- Roles: including end user, content expert of digital library project, system designer, and metadata team.
- Relations: including direct and indirect relationships. Direct relationships, which usually communicate through the metadata

team, exist between the metadata team, end users, content experts, and system designers. Indirect relationships exist between end users, content experts, and system designers.

Figure 3 The Metadata Service Model



3.2.10 Evaluation of metadata performance

The last stage of the MLM seeks to review results of the whole metadata process and performance. Depending on the collection projects' needs and focus, the evaluation can include the following aspects:

- The assessment of metadata record quality, including completeness, accuracy, record types, granularity levels, and records' serviceability. For example, an evaluation of the U.S. Government Information Locator Service (GILS) focused on assessing metadata quality from a set of criteria and procedures. (Moen, Stewart and McClure, 1998)
- The evaluation of the effectiveness of adopting a metadata scheme for retrieval, such as its capacity to facilitate access to title, creator, and subject.
- The evaluation of the use of metadata creation tools within the

collection project; for example, the Nordic DC Metadata Project performed the evaluation of the metadata creation tool and brought up valuable issues to be considered in the future design process. (Hansen, 2001)

- The investigation of how well each stage of the Metadata Lifecycle Model has been performed. Are there any stages that the collection project team feels need to be re-performed? Such as acquiring deep metadata needs.

4.0 Discussion and Findings: The NDAP Case Study

In order to learn how well the MLM has been used, a case study of over 40 collection projects from the Nation Digital Archives Program in Taiwan (NDAP) examined the feasibility of this approach. The collection projects of NDAP encompass various data types, disciplines, communities, and functions. The communities consist of libraries, archives museums, and herbariums. The disciplines comprise arts & humanities, social sciences, and natural sciences. The data types include photos, audio, films, rubbings, specimens, rare books, paintings, artifacts, documents, and dictionaries. They also request for diverse functions. The Metadata Architecture and Application Team (MAAT) aims to assist collection projects of the National Digital Archives Program (NDAP) in developing metadata mechanisms since 2000. How to construct a systematic approach to implementing metadata and content analysis for various digital libraries becomes an essential task to the MAAT. Under such complicated environments, the approach of MLM is introduced to achieve quality assurance, consistency, and interoperability. The case study is part of an on-going design process of the Metadata Lifecycle Model and hopes to learn more issues to be considered in the future development.

The questionnaires are sent by the MAAT to each collection project within NDAP in the beginning of every year to acquire supporting needs of metadata provision. Some collection projects hope to get support of the whole metadata lifecycle, because perhaps they do not have metadata specialists. Some collection projects need partial steps of the metadata lifecycle, for example, evaluation of relevant metadata standards, or assistance in preparation of metadata requirement specification. About 70% of the collection projects carry out the whole metadata lifecycle. The duration of the whole metadata lifecycle averages from four months to two years in this case. The findings resulting from the case study are described below.

4.1 Communication platform

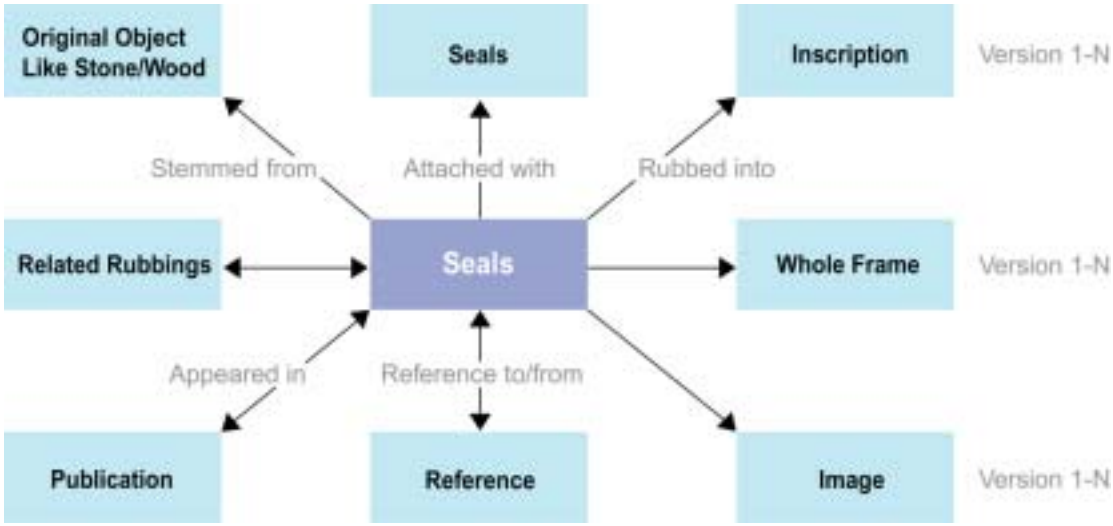
The collection project team, system development team, and MAAT all communicate metadata requirements through the MLM. Many collection project participants feel at ease, because the metadata system will be complete once all stages of the lifecycle are complete. It also relieves the burden of systems analysis from the system development team, as the MLM has brought much help toward this aspect. The MAAT feels it is a good method and efficient way of project management. The MLM helps the collection project team to express metadata requirements in a clear and systematic way, and plays a bridge between the collection project team and the system development team.

4.2 An association-rich approach for content-based analysis

A digital object may have rich associations with other objects due to its reproduction, digitization, and other activities. The content experts might need to identify both logical and physical relations for research. The end users might want to trace the related objects for more information. The Digital Archives of Chinese Rubbings is one example. A piece of Chinese rubbing is usually stemmed from a stone, wood, or bronze ware, and is enriched with a complicated range of relationships (as Figure 4). These relationships can be generalized into seven types:

- *Stemmed relationship:* a relation that indicates the original object for rubbings.
- *Attachment relationship:* a relation that labels that a seal is attached to a rubbing from collectors.
- *Form relationship:* a relation that explains that a rubbing can be produced into three forms, that is, full frames, images, and inscriptions.

Figure 4: A Context Graphic for Chinese Rubbings



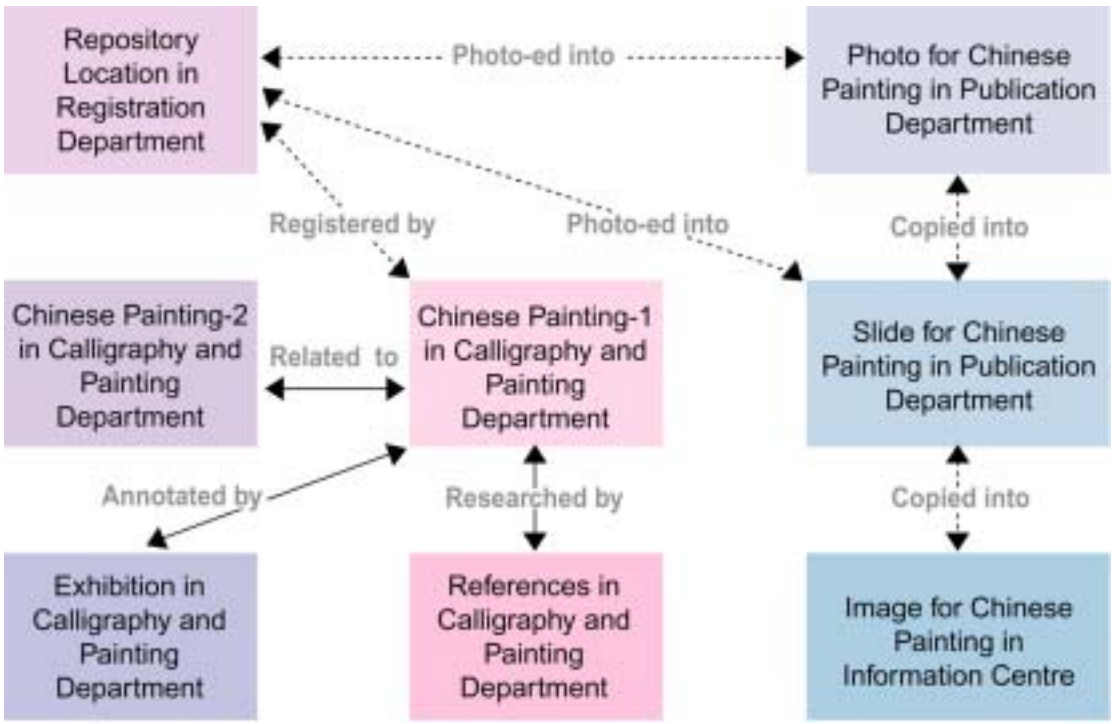
- *Version relationship:* a relation that describes that a rubbing for the same original object can be rubbed by different rubbing makers for the purpose of distinguishing one from the other based on quality requirement.
- *Related object's relationship:* a relation that illustrates that a rubbing is related to another one. For instance, the relation could be whole and part.
- *Reference relationship:* a relation that elucidates how many published papers research a specified rubbing.
- *Publication relationship:* a relation that reveals whether a rubbing is included into a specified publication.

4.3 A re-examination of workflow for the digital library

In order to illustrate and fulfill the metadata role and function, the related workflow and procedure should be re-examined. In the Digital Project of Chinese Calligraphy and Painting at the National Palace Museum in Taipei (NPM), Chinese paintings are managed and used by various departments for different purposes. From the metadata perspective, this is a multi-department collaboration project encompassing the Painting and Calligraphy Department, the Education and Exhibition Department, the Publication Department, the Registration Department, and the Information Center at the NPM. Though the metadata is mainly focused on arts research of Chinese paintings, the management requirements of different departments, such as digitization, exhibition, inventory, preservation, resource discovery, and rights management, need to be included into metadata construction and analysis. Therefore, existing workflow and procedure (as in Figure 5) should be

re-examined and re-engineered in a thorough analysis to define the roles and functions for metadata requirements.

Figure 5: A re-examination of workflow for the National Palace Museum



4.4 Granularity issue

With the process of digitization, the content of original objects could be freed from their previous media and have the chance to divide into different levels of granularity. These objects later can assemble into larger structures to facilitate not only retrieval and reuse but also recreation for future repurpose. The granularity refers to how finely chopped is the metadata (Tennant, 2002). In the case of Institutional Project of Taiwan Historica, the EAD schema has been used to elucidate hierarchical levels and associations for the archive collection. Therefore, the information granularity is separated into seven levels from top to bottom: fonts, sub-fonts, series, sub-series, file, item, and attachment.

4.5 More than one metadata scheme

A collection project may adopt different metadata standards according to its different functional requirements. For example, The Rare Books Digital Archives Project adopts MARC21, EAD, and TEI as the metadata schemes. The MARC21 serves to support the descriptive function. The EAD enables browsing the collection. The TEI serves to represent the textual features and tagged structures needed for research, such as individual poems in the

collection used for online research. In some cases, the introduction of MLM gives the collection project more appropriate ways to structure information. For example, the Digital Library Project of Exploration of Taiwan Visual Image included the VRA3.0 (Visual Resources Association) and IFLA FRBR (International Federation of Library Associations and Institutions, Functional Requirements for Bibliographic Records) Model as the metadata schemes, after the processes of investigation of deep metadata needs and identification of strategies for the metadata schemes within the MLM completed. In addition, the collection project needs other kinds of metadata attributes, such as person, place name, time-span, and event.

4.6 Different Dimensions of Metadata Sets

For many digital libraries, discovery, access, and description are the typical applications of metadata. However, the case study shows that there are more functions needed, such as preservation, content rating, rights, data management, and e-Learning. In addition, domain-specific metadata sets are the core development of metadata practices in the case study. In other words, the generic metadata set, such as Dublin Core for resource discovery, is deployed in its top level of the union catalogue. The domain-specific metadata sets, such as CDWA and EAD, are used in the advanced setting of the union catalogue within the case study.

5.0 Conclusion and Future Work

Digital libraries, through the use of metadata systems, have great potential to benefit society as bearers of valuable information resources. The Metadata Lifecycle Model (MLM), an evidence-based approach that normalizes metadata requirements, has been proved to achieve efficiency, quality assurance, and consistency for the NDAP by providing a systematic way to develop metadata systems. It is our hope that the MLM will have the opportunity to similarly assist other digital library projects by providing them valuable, evidence-based guidance.

In the near future, the issue of interoperability will become one of the urgent metadata tasks for the NDAP case. To this day, the case has accomplished a series of common core element sets in order to achieve discipline-based information sharing for various communities composed of a wide range of projects, such as Chinese rare books, botany, zoology, and archives. In 2003, several common core element sets will be completed based on results from requirement surveys by

questionnaire. In parallel, it is necessary to build up a centralized bilingual metadata registry consisting of both Chinese and English for metadata interoperability and promotion of data standard usage, which will consist of metadata standards, application profiles, best practices, specifications, controlled vocabularies, and database schemas. Furthermore, a model of metadata standard selection will also be developed to achieve greater effectiveness of metadata implementation.

Reference

Gilliland-Swetland, A. (1998). Setting the Stage. In M. Baca (Eds.), *Introduction to Metadata: Pathways to Digital Information*. Los Angeles, Calif. : Getty Information Institute.

Hansen, P. (2001). Evaluation and design issues of Nordic DC metadata creation tool. In *Proc. Int'l. Conf. On Dublin Core and Metadata Applications 2001*. Japan: National Institute of Informatics [Online]. Available:

Hudgins, J., Agnew, G., & Brown, E. (1999). *Getting mileage out of metadata: Applications for the library.* Chicago: American Library Association.

Moen, W.E.,Steward, E.L., & McClure, C.R. (1998). Assessing metadata quality: Findings and methodological considerations from an evaluation of the US government information locator service. In *IEEE International Forum on Research and Technology Advances in Digital Libraries : ADL 98 proceedings* (pp. 246-255). Los Alamitos, CA: IEEE Computer Society Press.

Tennant, R. (2002). Digital libraries: The importance of being granular. *Library Journal, 127*(9), 32-33.

Zeng, M. L. (1999). The Metadata elements for object description and representation: A case report from a digitized historical fashion collection project. Annual Review of OCLC Research 1999 [Online]. Available: http://www.oclc.org/research/publications/arr/1999/zeng/