

A Metadata Lifecycle for Content Analysis in Digital Libraries

Ya-ning Chen Shu-jiun Chen Yi-ting Chang Simon C. Lin

**Computing Centre, Academia Sinica
Taipei, Taiwan, R.O.C.**

A Metadata Lifecycle for Content Analysis in Digital Libraries

Ya-ning Chen

Systems Analyst

arthur@sinica.edu.tw

Shu-jiun Chen

Systems Manager

sophy@sinica.edu.tw

Yi-ting Chang

Information Staff

sat@gate.sinica.edu.tw

Simon C. Lin

Director

sclin@sinica.edu.tw

Computing Centre, Academia Sinica

Taipei, Taiwan, R.O.C.

摘要

在網際網路領域中，資訊無序已是一項日趨嚴重的問題，如何賦與資訊結構化，以找出精確且質佳的資訊，也成為全球數位圖書館計畫中，極為重要的研究課題之一。針對國家數位典藏計畫的需求，中央研究院 Metadata 工作組研發一套 Metadata 生命週期(metadata lifecycle)，以系統化方式進行內涵分析、Metadata 的應用，以及計畫管理等多層功能。Metadata 生命週期共由八個基本單元組成：「需求訪談」、「主題計畫的需求與屬性分析」、「主題計畫相關標準與趨勢的剖分」、「資訊需求分析」（包括 Metadata 與主題計畫元素的核對、元素分佈、類型與範圍、不同系統資料庫的關聯、國際標準的映對、元素的定義與範例）、「Metadata 需求規格書的制訂」、「Metadata 系統工具的評估」、「規範指引的製作」、「發展 Metadata 的基準試驗平臺(test-bed)」，以及「Metadata 服務」等。本文發現此一 Metadata 生命週期應用在各項主題計畫中，產生極為具體的成效，包括「Metadata 形態與元素的分佈」、「富關聯性的內涵分析」、「資料庫關係的分割與整合」、「作業流程

的檢視」、「平行式國際標準的接軌」。

Abstract

The issue of chaos order in digital information on an Internet scale has been recently raised for many digital projects around the world. Metadata is an emerging approach to improve precision for resource discovery. The aim of this paper is to present a metadata lifecycle with nine components as the basic model to support content analysis and organize digital information within structured and associated context for the digital library. The lifecycle has shown useful to build up several benefits in terms of metadata process for the digital library programme. The benefits include an analytical distribution of metadata types and elements, a relationship-rich approach for content analysis, a context-centric analysis for system integration, a re-examination of workflow, and a two-parallel orientation to metadata standardization.

Introduction;

Metadata is an emerging approach to organize digital information in a structured manner and support precise retrieval for digital libraries on an extraordinary Internet scale. Although there are many metadata practices in digital libraries, few literatures have been noted about how to choose the right metadata formats for their own projects. This paper aims to introduce a metadata lifecycle developed by the Academia Sinica as a basis of content analysis to serve the functions of choosing right metadata standards for the National Digital Archive Initiative sponsored by the National Science Council in Taiwan. More than ten projects of the Initiative are employed as the case study to elucidate the framework of metadata lifecycle and show the findings. The issue of metadata system design and implementation is related to content of the metadata lifecycle, but is not addressed in this paper. The metadata lifecycle consists of nine parts as follows: interview with content experts, analysis of project requirements and attributes, review of relevant metadata standards and projects, analysis of information requirements, preparation of the metadata requirement specification, evaluation of metadata system and development, preparation of best practice, development of metadata test-bed, and maintenance of metadata service.

Definition

Some relevant literatures have offered definitions of content analysis. Based on Bos, & Tarnai's point of view, content analysis is a means of analyzing texts (Bos, & Tarnai, 1999, p. 660). The Writing Center at Colorado State University regarded

content analysis as a research tool used to determine that presence of certain words of concepts within texts or sets of texts (The Writing Center, n.d.). From Stanton's conceptual perspective, content analysis is thought as a technique that has been around since the beginning of the century of analyzing the content of documents. The term "document" refers to all media: newspaper, diaries, speeches, letters, reports, books, journals, notices, films, photographs, videos, radio, and television programmes (Stanton, 1995, p. 7/2). However, content analysis is a research tool or technique deployed to clarify the content of document for various purposes.

A Metadata Lifecycle

In an era of digital libraries, metadata is often used to organize information in an order way to support a better resources discovery and retrieval. It is very important to understand content of document prior to applying any specific metadata formats or standards for the digital libraries, so content analysis is essentially required for any digital library projects. According to Stanton's concept, content analysis can be divided into 5 stages as follows: determine objectives, define unit of analysis, construct categories for analysis, test coding to assess reliability, and conduct analysis (Stanton, 1995, p. 7/2-7/3). Stanton's conceptual ideas are mainly focused on document analysis for designing a hypermedia system, so it is a kind of computer system approach to analyze document. On the other hand, Hudgins, Agnew, & Brown planned 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, maximize 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, reporting results, and conclusion (Hudgins, Agnew, & Brown, 1999, pp. 42-53).

Over 20 projects demand for metadata plan and implementation in the Digital Archive Initiative supported by the National Science Council in Taiwan since 2000. In order to achieve a consistent structure for these projects, a metadata lifecycle is designed by the Sinica Metadata Architecture and Research Taskforce (SMART) for this requirement in terms of both project management and content analysis. The metadata lifecycle is composed of nine components and can be triggered once again while a new or change of project requirements for metadata is initiating. These tasks for the lifecycle are conducted by a series of questionnaires and tables. However, these components for the metadata lifecycle are composed by follows: interview with content experts, analysis of project requirements and attributes, review of relevant metadata standards and projects, analysis of information requirements, preparation of

the metadata requirement specification, evaluation of metadata system and development, preparation of best practice, development of metadata test-bed, and maintenance of metadata service.

Interview with content experts

The first step of the metadata lifecycle is to take a face-to-face interview with content experts and to get an overview of their metadata requirements for each content project. Prior to interview session, two tasks are necessary to undertake. First, members of the SMART group have to take a serious examination of project background information based on review of project proposal such as purposes, goals and expected results. Second, the SMART group sends questionnaires to content projects and inquires information about scope, metadata element and structure, legacy record and system, metadata context, expected result for different stages, contact information, and so on. 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 metadata are expected to accomplished?
- ***Metadata scope:*** how many types of metadata are required for the projects? Such as types of object, person, event, temporal terms control and expression, and geographic name.
- ***Legacy record and system:*** basic information about the learning system, including metadata elements, structure, and number of records, storage format, input method and system. In addition, it is useful to understand the advantage and disadvantage of the legacy system.
- ***Metadata context:*** Is only one metadata database constructed for this project? Are any other databases required to integrate with this metadata database, like 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 content analysis?

Analysis of Project Requirements and Attributes

A task of comprehensive analysis is required to ensure requirements and attributes of metadata. The project requirements of metadata would be verified in a systematical way after a detailed discussion in interview session. Several agreements should be defined clearly and attained to prepare the related tasks at next session, like metadata schedule, scope, context, and role and function. Certainly, some real examples should be collected to help the SMART members in understanding project goals and

meanings for each element.

Review of relevant Metadata standards and projects

The most important metadata task is to select an appropriate metadata standard, instead of developing a new standard. In this session, the SMART group takes a serious examination and survey of existing metadata standards and relevant projects, and then offers a comprehensive comparison between standards and metadata requirements of each project to achieve few results. First, current trends and issues related to the implementation of metadata standards or projects around the world can be discovered and provided an advice for practical application and future development. Second, project members could well know what kind of differences is from other similar or homogeneous projects at the same time, and re-arrange the focus of expected project goals. Third, project's objectives could be segmented into different parts, and the right metadata standard can also be decided.

Analysis of Information Requirements

In preparation for analysis of information requirements, several works should be undertaken as a basis for analyzing and ascertaining the project requirements. First the definitions and examples of initial metadata elements should be offered and would be clarified after interview and adjustment. Second, the proposed metadata standard and elements list are selected and explained for project participants. Third, the comparison among selected standards and required elements of the project is conducted. Fourth, an analytical context diagram with various relationships for metadata scope and context would be defined. Fifth, indexing keys and access points would be advised as a basis for system design, as well as for metadata role and function. In this session, the benefits are as follows:

- Metadata elements and categories are chosen and defined clearly based on a comparison with an existing metadata standard.
- Distributions of metadata elements are verified in compared with selected metadata standards. These include the distribution of description, administration, system management, and rights management, resources discovery.
- Metadata scope and context are clarified, and related relationships are also drawn a clear line and attributed to a diversity of categories.
- It could ensure what kind of systems and databases are integrated by metadata mechanism such as GIS.
- A crosswalking is accomplished between existing metadata standards and project required metadata elements.
- Real examples and definitions for projects are collected as a basis for the best

practice.

Preparation of the Metadata Requirement Specification (MRS)

To achieve a common agreement among project participants, metadata members and system designers, a Metadata Requirement Specification (MRS) is prepared by the SMART group as a bridge for these members across a variety of disciplinary domains. The specification contains the several components: executive summary, background information of project, project participants, attributes of metadata elements (like label, length, type (like alphabetical, numeric, or both), index key, and so forth), an input template, a mapping table between metadata standards and project requirements, a context diagram for metadata scope among a range of systems, XML DTD, 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 for project members to check against with metadata system and function.
- A basis in preparation for a best practice and a crosswalk mechanism among existing metadata standards.

Evaluation of Metadata System

A metadata system is a tool to present the concrete result after a series of analyses and preparations of specification. From system evaluation perspective four criteria developed by the Federal Geographic Data Committee are useful to chose or develop metadata systems (as Table I). System evaluation is an essential task for the SMART group and its supported projects, because this task provides the basic reference:

- To ensure what kind of metadata systems and tools are available for existed metadata standards.
- To learn what kind of functions can be achieved, and what kind of unsolved issues are encountered for the specific metadata standard.
- To know that future trends and developments are proposed for the specified metadata standard.
- To ascertain what kind of customized tasks are required for a Chinese version.

- Metadata exchange: including import capabilities, export capabilities, and completeness/compliance.
- Usability: including user interface, user prompting, minimizing duplication of entry, data generation integration, cut and paste, restart-ability, documentation, and miscellaneous.
- Administrative: including platforms/installation, stand-alone, updates.

- Tool reliability: including robustness, recovery mechanisms, and consistency.

Table I: Evaluation Criteria for Metadata Tools
(Source from: The Federal Geographic Data Committee, n.d.).

Preparation of Best Practice

A best practice is often created after completing metadata analysis and system. The term “best practice” is commonly used in a wide variety of domains for different purposes. It could be deployed as a document to reach six basic functions: introductory, definitive, explanatory, instructive, paradigmatic, and standardized. Several definitions for best practice have been elucidated as follows:

- A best practice is one of many ways of documenting and sharing problems and solutions related to improving the health of patients and communities (The Best Practice Network, n.d.).
- The best practice documents bring together practical guides as well as background documentation and data that support the recommendations and standards that have made the project so successful (Quam, 2000).
- The code will assume that, taking into account nationally agreed principles and practices each institution has its own systems for independent verification both of its quality and standards and of the effectiveness of its quality assurance systems (The Quality Assurance Agency for Higher Education, n.d.).
- A best practice is comprised of policies, principles, standards, guidelines, and procedures that contribute to the highest, most resource-effective performance of a discipline (Finneran, 2001, p. 3).

In the Academia Sinica, the best practice of metadata for any digital archive projects is composed by following components: a briefing of project background, principles for designing a metadata set and related elements, field’s label, scope and definition, instruction, example, related standards and their applications, suggestion for system implementation, and so forth. However, consistency and quality assurance are two most important objectives as planned to accomplish. The best practice is not only employed for a quality control mechanism, also regarded as the useful document for communication of all project members, transferring from a legacy system to new one, and information sharing with other projects.

Development of Metadata test-bed

At present two options are used by the SMART team in the Academia Sinica. One is to select an existing system developed by the homogeneous or similar projects in domain of digital library. In general, this option is deployed as a prototype for each

project in the Academia Sinica to justify the rightness of metadata elements. Further, also employed as an objective feedback for adjusting the metadata elements, and for revising the metadata specification. If the test-bed system is fit for project requirements, then the system is required to customize into a Chinese mode. On the other hand, one generic system is under development by the Academia Sinica for about 20 projects of the Digital Archive Initiative, and is proposed to attain a kind of system integration and interoperability. Furthermore, the generic system will be a key component of metadata clearinghouse across more than 20 projects in the near future for various purposes, like a federated meta-search engine.

Maintenance of Metadata Service

In order to undertake the metadata and content analysis for projects that the SMART supports, several fundamental metadata services are developed to guarantee the quality assurance. A service model is constructed to formulate the service items. The metadata service model is composed by four basic elements as follows (refer to Figure I): role, relation, service content or responsibility, and service mechanism.

- ***Role:*** including end user, content expert of digital library project, metadata system designer, and metadata team.
- ***Relation:*** including direct and indirect relationships. The direct relationships can be categorized between metadata team, end user, content expert, and system designer. The indirect relationships can be categorized between end user, content expert, and system designer.
- ***Service content or responsibility:*** including a wide diversity of knowledge assistances in:
 - ▲ *Designing* a user interface and related function for end user.
 - ▲ *Providing* consultation in delivering knowledge of metadata and content analysis to content expert.
 - ▲ *Implementation and construction* of metadata and related standards for content expert.
 - ▲ *Developing* a best practice for content expert.
 - ▲ *Accomplishing* a crosswalk mapping among legacy records, project required metadata fields, and selected metadata standards for system designer.
 - ▲ *Designing* metadata system and interface for system designer.
 - ▲ *Offering* advice on an issue of interoperability across a wide range of metadata standards, like crosswalk, meta-search, and so on.
- ***Service mechanism:*** a range of mechanisms supported to undertake metadata tasks in a background context
 - ▲ *Content analysis and relationship:* including object-oriented approach, IFLA's

functional requirements for bibliographic records (FRBR) model, and a core-individual structure for metadata record.

- ▲ *Crosswalking*: besides Dublin Core is a default mapping for any projects, another domain-specific metadata mapping is also offered such as CDWA, EAD, TEI Lite, FGDC, and so on.
- ▲ *Evaluation*: including cost, time, procedure, service, human resource, quality assurance and consistency, and system functionality for each content project management regarding to metadata implementation and construction.
- ▲ *Clearinghouse*: is under development for information sharing like online mapping and meta-search for different metadata standards, and so on.

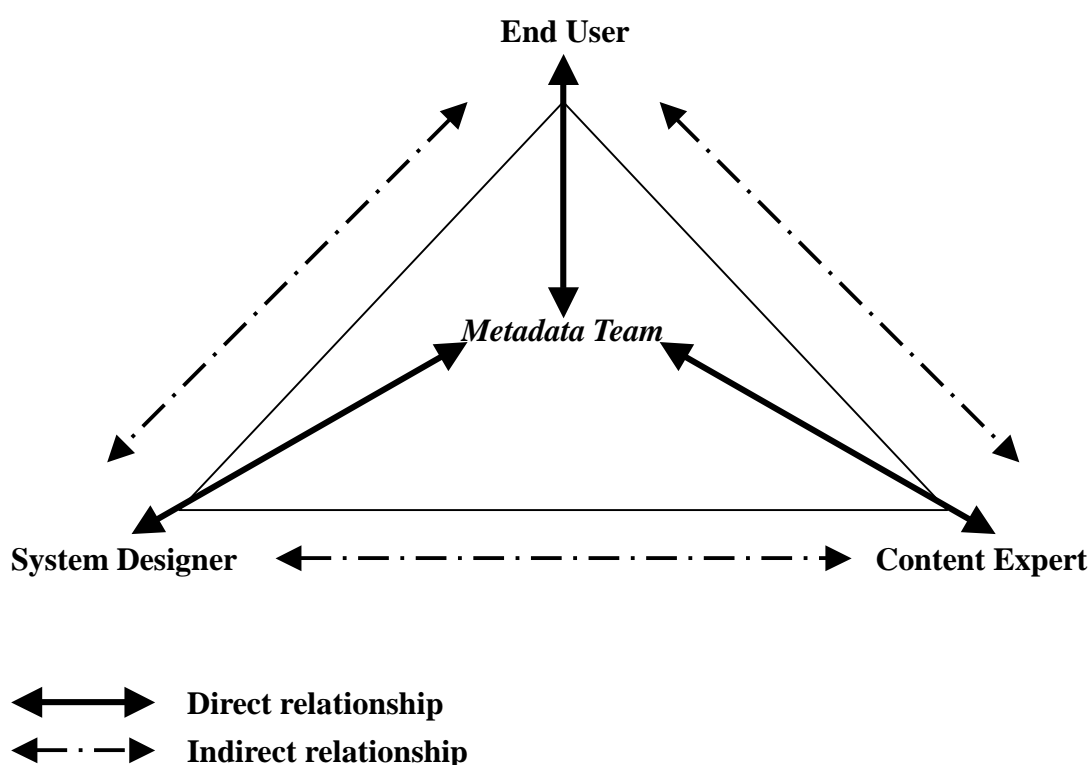


Figure I: A Metadata Service Model

Case study and related findings

In order to learn how the metadata lifecycle is used in the Academia Sinica, several pilot projects from the Digital Archive Initiative are selected to clarify content of the metadata lifecycle with examples. Currently the SMART group has supported about 20 projects for metadata requirements. These pilot projects have mainly been commenced by the Academia Sinica since 2000 and start to enforce in 2001. They cover attributes across various domains and request for diverse functions, and attributes of projects can be categorized into three types (as Table II). How to

construct a systematic approach to implementing metadata and content analysis for various projects becomes an essential task for the SMART group. Under such complicated environments, the metadata lifecycle is formulated to achieve quality assurance, consistency, and interoperability. After one year's practical experiment and implementation, the metadata lifecycle brings advantageous as follows:

- **Organization:** including museum, archive, herbarium, and library.
- **Subject domain:** including ethnology, history, arts, geography, archaeology, linguistics, bio-diversity, and the Chinese literature.
- **Data type and style:** including Chinese rubbings, rare books, archives, corpus, specimen, artistic works, and field works.

Table II: Attributes of pilot projects supported by the SMART team

An analytical distribution of metadata types and elements

Based on implementation of metadata for different projects, the SMART group finds that metadata types are generalized into 5 types: object, person, event, temporal, and geographic/place name. Most of existing metadata standards and elements are focused on object, but few are not. For instance, content standards for digital geo-spatial metadata (CSDGM) developed by the Federal Geographic Data Committee (FGDC) in USA is a spatial emphasis on geographic metadata information. Within 20 supportive projects in the Academia Sinica, the Taiwan Aborigine Pin-pu Group Project is an ethnology project of aborigine group research. The SMART members offer one distribution of metadata types and elements as an analytical reference (temporal: 4, spatial: 4, person: 11, and event: 4 elements) for project members after interview and analysis of information requirements. Indeed, this reference reveals a hint evidence for this project. In effect metadata types and elements for this project should be mainly scattered about persons or peoples, and geographic metadata, but the first initial consequence is not matched with project's attributes based on the analytical reference. An advice for re-arrangement of distribution of metadata types and elements is offered to the content project, and a balance agreement is achieved to turning focal points into geographic metadata information.

A relationship-rich approach for content analysis

Another focus of content analysis within the metadata lifecycle is both of logical and physical relationships for research purpose. The Academia Sinica is the major stakeholder of Chinese rubbings, though different organizations across Japan, France, USA, and Mainland China also house. In the Academia Sinica, Chinese rubbings are

categorized into four types: Buddhism Images, Stone Tablet Texts (inscriptions), Tomb Tablet Texts, and others. The Stone Tablet Texts are the key collections among these four. Generally, research domain of the Chinese Rubbings Project is across arts, history, and the social sciences. In essence, a piece of rubbing is stemmed from a stone, wood, or bronze ware, and is enriched with a complicated range of relationships (as Figure II). These relationships can be generalized into seven types:

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

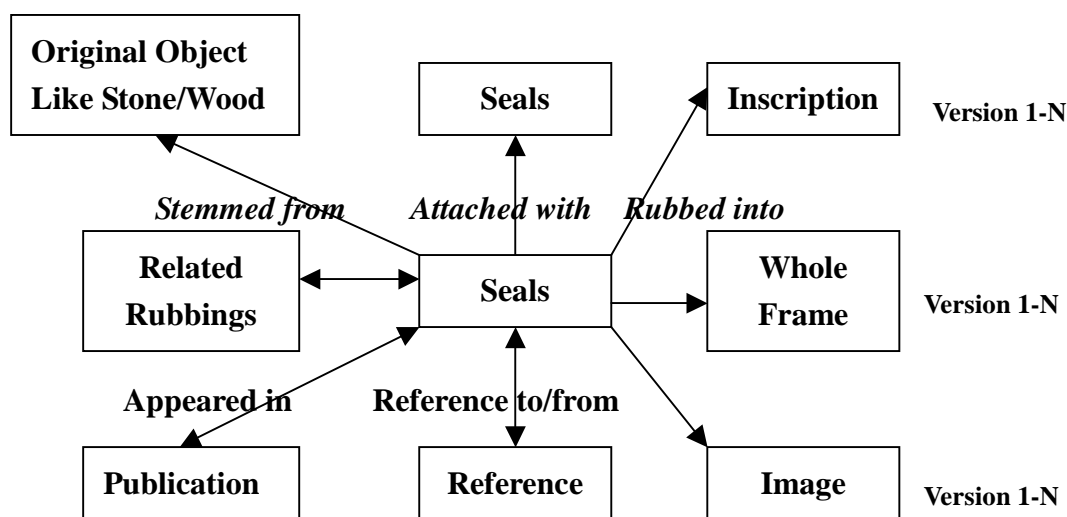


Figure II: A Context Graphic for Chinese Rubbings

- **Version relationship:** a relation 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 illustrates that a rubbing is related to another one. For instance, the relation could be whole and part.
- **Reference relationship:** a relation elucidates that how many published papers are researched for a specified rubbing.
- **Publication relationship:** a relation reveals whether a rubbing is included into a specified publication.

A context-centric analysis for system integration

A context analysis is served as a basic reference for system integration. In the

Zoological Research Project of Fish in Taiwan, the project covers three dimension facets, that is, species, gene, and ecology, as well as related specimen in bio-diversity domain. After an analysis of project requirements, a context graphic with relationships (as Figure III) between different metadata systems is illustrated clearly. In terms of system integration and connection, the context graphic is very helpful for project to decide three matters. First relates to the project schedule and expected results. In the Zoological Research Project of Fish in Taiwan, the project has selected species as the first priority goal. Second, the content project can choose the workable scope and objectives, and then segment project schedule and expected results into annually for project management. Third, project also learns well that how to establish a connection and association related to other bio-diversity systems around the world.

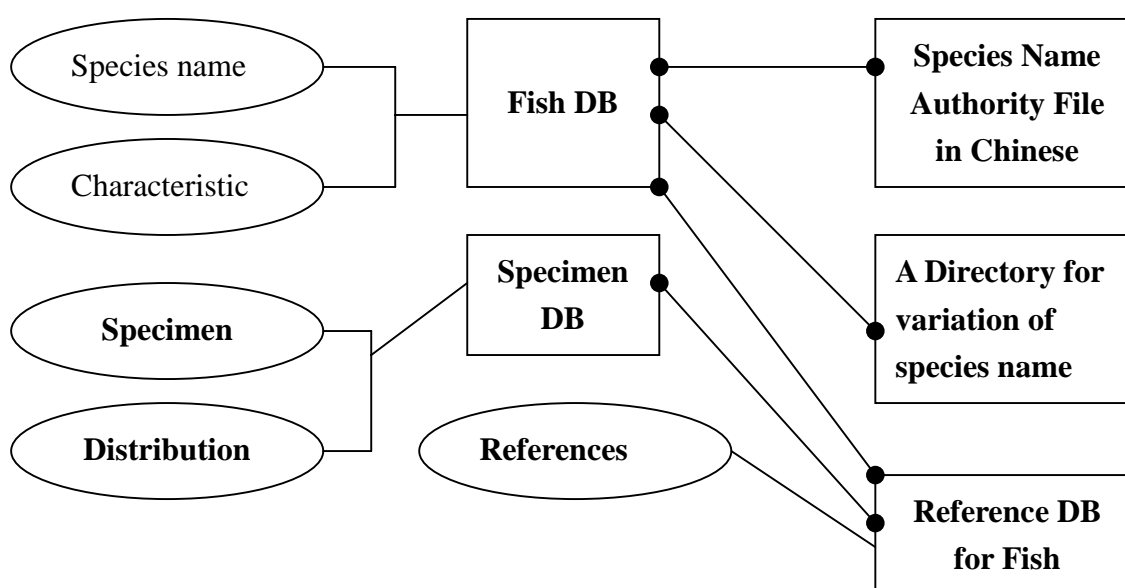


Figure III: A Context Graphic for System Integration

A re-examination of workflow

In order to illustrate and fulfill the metadata role and function, related workflow and procedure should be re-examined again. 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 among 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 are 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, existed workflow and procedure (as Figure IV) should be re-examined and re-arranged in a serious analysis

to define the roles and functions for metadata requirements.

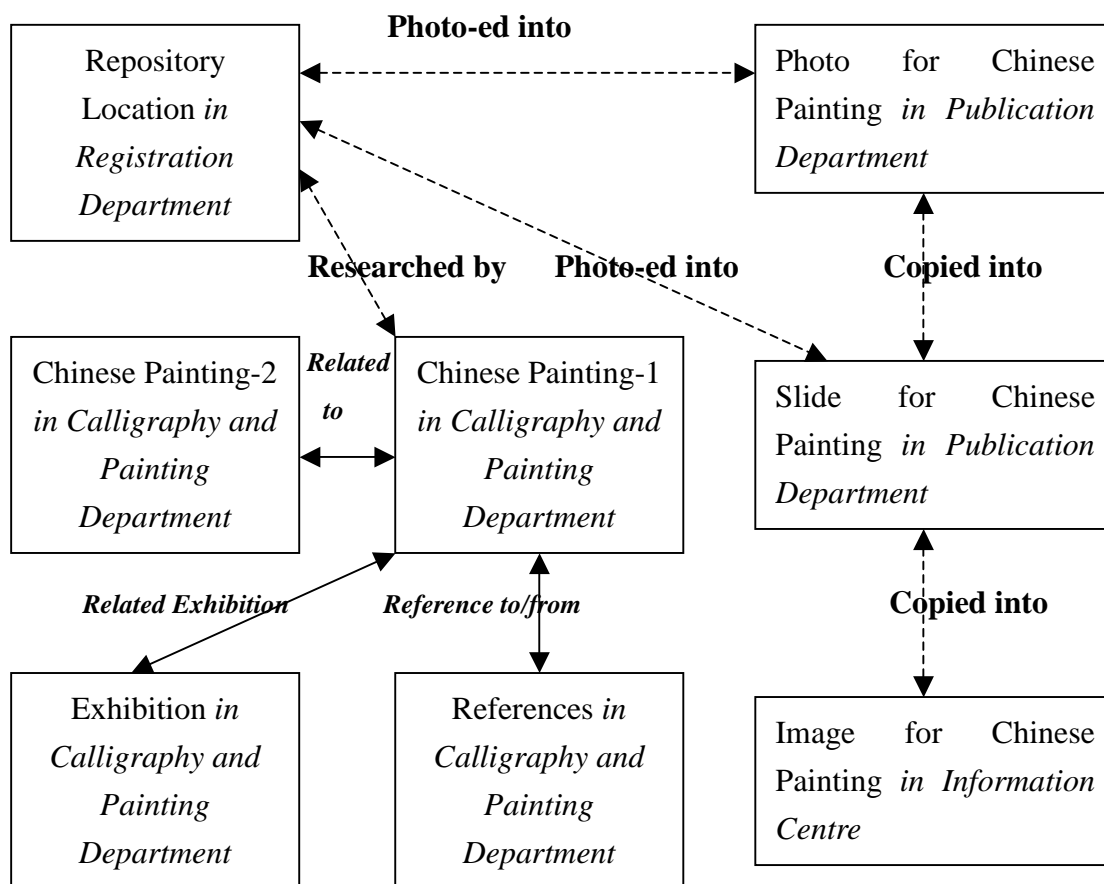


Figure IV: A re-examination of workflow for the National Palace Museum

A two-parallel orientation to metadata standardization

At present one principle adopted by the SMART for any digital projects is to employ the Dublin Core and another one domain-specific standard at the same time for the purposes of standardization and crosswalk. In essence, the Dublin Core Elements Set is very simple and popular. It is also a highly conceptual metadata set served as a common crosswalk ground for mapping and federated meta-search. On the other hand, we can also find that the Dublin Core is too ambiguous for most of research-focused digital archive projects supported by the SMART. However, a principle of adopting both of the Dublin Core and another specific comprehensive metadata standard simultaneously is deployed. The two-parallel principle to standardization can bring two obvious benefits:

- A common interface with a federated meta-search engine across various metadata systems can be developed for a wider range of digital projects.

- A precise crosswalk between different metadata formats can be achieved.

Conclusion

It is no doubt that the metadata lifecycle developed by the Academia Sinica is very advantageous for project management, and brings many benefits for implementation and construction of metadata in terms of content analysis. The metadata lifecycle is also served as a collaborative ground among content expert, system designer, and metadata members to achieve expected roles and functions for metadata in digital library domain. Further, the lifecycle can be regarded as a useful basis for evaluation on cost, time, human resource, quality assurance, and system function in light of project management. However, several fundamental issues in information granularity, a crosswalk mechanism for legacy systems and records, and a clearinghouse mechanism are still under development, and will be addressed in another paper in the near future.

Reference:

- The Best Practice Network. (n.d.). Preparing your best practice submission.
<http://www.best4health.org/solutions/bestpractices/index.cfm> (visited 20 September 2001)
- Bos, W., & Tarnai, C. (1999). Content analysis in empirical social research. *International Journal of Educational Research*, 31, 659-671.
- The Federal Geographic Data Committee. (n.d.). Evaluation criteria.
<http://www.fgdc.gov/clearinghouse/mitre/task2/evalcriteria.html> (visited 21 September 2001)
- Finneran, T. (2001). A best practices assessment.
<http://www.ciber.com/downloads/whitepapers/bestpractice/ciberbestpractices.pdf>
(visited 20 September 2001)
- Hudgins, J., Agnew, G., & Brown, E. (1999). *Getting mileage out of metadata: Applications for the library*. Chicago: American Library Association.
- The Quality Assurance Agency for Higher Education. (n.d.). Code of practice for the assurance of academic quality and standards in higher education.
<http://www.qaa.ac.uk/public/COP/codesofpractice.htm> (visited 20 September 2001)
- Quam, E. (2000). State of Minnesota best practice guidelines for web metadata.

<http://www.bridges.state.mn.us/bestprac> (visited 20 September 2001)

Stanton, N. (1995). Content analysis: A methodology for hypermedia design. Paper presented at The IEE Colloquium on Authoring and Application of Hypermedia-based User-interface.

The Writer Center at Colorado State University. (n.d.). Content analysis.

<http://writing.colostate.edu/references/research/content/index.htm> (visited 12 September 2001)