The Design of Metadata Schemas for Chinese Cultural Celebrities' Manuscript Library

Wei Liu
Email: wliu@libnet.sh.cn
(Digital Library Institute of Shanghai Library, Shanghai, China 200031)
Cuijuan Xia
Email: cjxia@libnet.sh.cn
(East China Normal University, Shanghai, China 200062)
Liang Zhao Xiangying Lou Chunjing Zhang
Email: cjzhang@libnet.sh.cn
(Computer and Network Center of Shanghai Library, China 200031)

Abstract: Metadata is essential to the semantic discovery of a digital library. A lot of metadata standards have been issued in recent years, but few standards or guidelines are actually implemented. Practitioners are always suffering from no guidelines or pilot projects can be referenced when developing digital library applications. The situation does no good to overcome interoperability problems between applications. This paper summarized the design of metadata schemas for Chinese Cultural Celebrities' Manuscript Library (CCCML), which is a branch of Shanghai Library. In practicing a formalized and normalized way to implement the merits of metadata for CCCML, and overcome the confusion of using metadata, we experienced an approach in deriving a set of principles, delivering an outline of methodology, conducting a procedure for the practice, and developing the specification of the schemas, as well as the implementation of the schemas in the CCCML digital library system.

Keywords: metadata, metadata application, digital library, manuscripts, Dublin Core, Shanghai Library

1. Introduction

In recent years, the word "metadata" has become very popular not only in the field of computer science but also in the field of knowledge management and library circle. Efforts all over the world have been made to standardize metadata. Especially when the web technology involved with semantics, metadata becomes the most significant building blocks to the semantic web architecture. Up to now, dozens of metadata formats, such as MARC, TEI, CDWA, EAD, DC, and etc. are widely used in hundreds of applications. And there are more and more metadata formats to be introduced in domain specific applications. Among which Dublin Core (DC) is always considered the most generalized and simplified format that is very powerful for its interoperability. DC's fifteen elements always act as a CORE set of metadata in a large number of digital library applications. Obviously, just the core is far from sufficient to almost any application. Rules and specifications to qualify or extend the core need to be developed. Dublin Core Metadata Initiative (DCMI), which is the host of the DC metadata, recommends the "Metadata Application Profile" (MAP) as a way to implement the DC metadata and as well as to fulfill the local metadata needs.

This paper is a summary of the design of a set of DC-based metadata schemas based on MAP for CCCML (Chinese Cultural Celebrities' Manuscript Library) digital library system.

2. Background

As one of the most influential library in China, Shanghai Library holds abundant repositories. Since 1996 Shanghai Library has devoted a lot of resources to build and manage the Chinese Cultural Celebrities Manuscript Collection. CCCMC is one of the unique and precious repositories of Shanghai library. No later than the end of 2005, Chinese Cultural Celebrities Manuscript Library will be moved to a renovated historical building located in Bund area, which is the most famous cultural district in Shanghai. This digital library system is especially developed to showcase its collections.

A prototype of the digital library system is scheduled to be constructed by the end of 2004. In the first stage, the digital library system not only supports the basic metadata retrieval functions and inventory management, but also showcases the diverse types of its collections and complicated linkage to various databases and systems within the library and as well as outside of the library. About one tenth of its collections (more than 5000 pieces) will be digitized at this stage. A group of people from Technical Service Department and Historical Collection Department joined the efforts in the requirements analysis and system development. And the coding will be outsourced to a software company in Shanghai.

The metadata schemas should be the first consideration to start the whole project. It's

https://doi.org/10.23106/dcmi.952107854

essential for the requirements analysis and system design. Metadata brings the structure to the data, and provides the base architecture to the system. The metadata profile for CCCML should also provide a mechanism or container for all the needed properties required by all kinds of types of resources and as well as semantic restrictions and encoding constrains.

The working group of the project surveyed and reviewed the related metadata standards, specifications and projects worldwide after the project was kicked off. One project, MALVINE[1] (Manuscripts and Letters via Integrated Networks in Europe) came into the sight with it's similarity in the aspects of attribute description requirements and technical environment. The working group of CCCML got the list of metadata elements of MALVINE project after contacting them. MALVINE's list is proven to be very helpful for deriving the metadata elements for CCCML project.

3. Requirements

The collections of CCCML, which contain various formats and types of diversified literal and cultural relics, are very diferent from the collections in traditional libraries. Currently, there are about a dozen of collections in CCCML. They are:

- Manuscripts
- Letters
- Diaries
- Photographs
- Paintings and calligraphies
- Books with signature and remarks
- Print materials
- Notebooks
- Audio and video (AV) Materials
- Certificates
- Diploma
- Physical objects

These categories are decided by the content expert of the library, which is far beyond exhausted. The categories may increase as the collections grow and new "discoveries" emerge from the inventory; therefore the system should be flexible enough to support dynamic formation of multiple schemas. First, it needs to clarify the relations among the properties of every type of resources and the relations among the resources themselves. See figure 1 for the ER model of resources in CCCML.

Each resource of the collections has to be defined unambiguously. This can help to emboss and differentiate resources; this also makes it easier to keep the consistency for choosing the common properties of all types of resources, finally it helps to describe the relations between the types of resources. One specification manual called "Resource Analysis Report" is introduced to define the resource, clarify its scope and boundary, establish the relationships between various types of resources or objects, decide the granularity of the object to be described, and finally confirm the retrieval and access requirements.

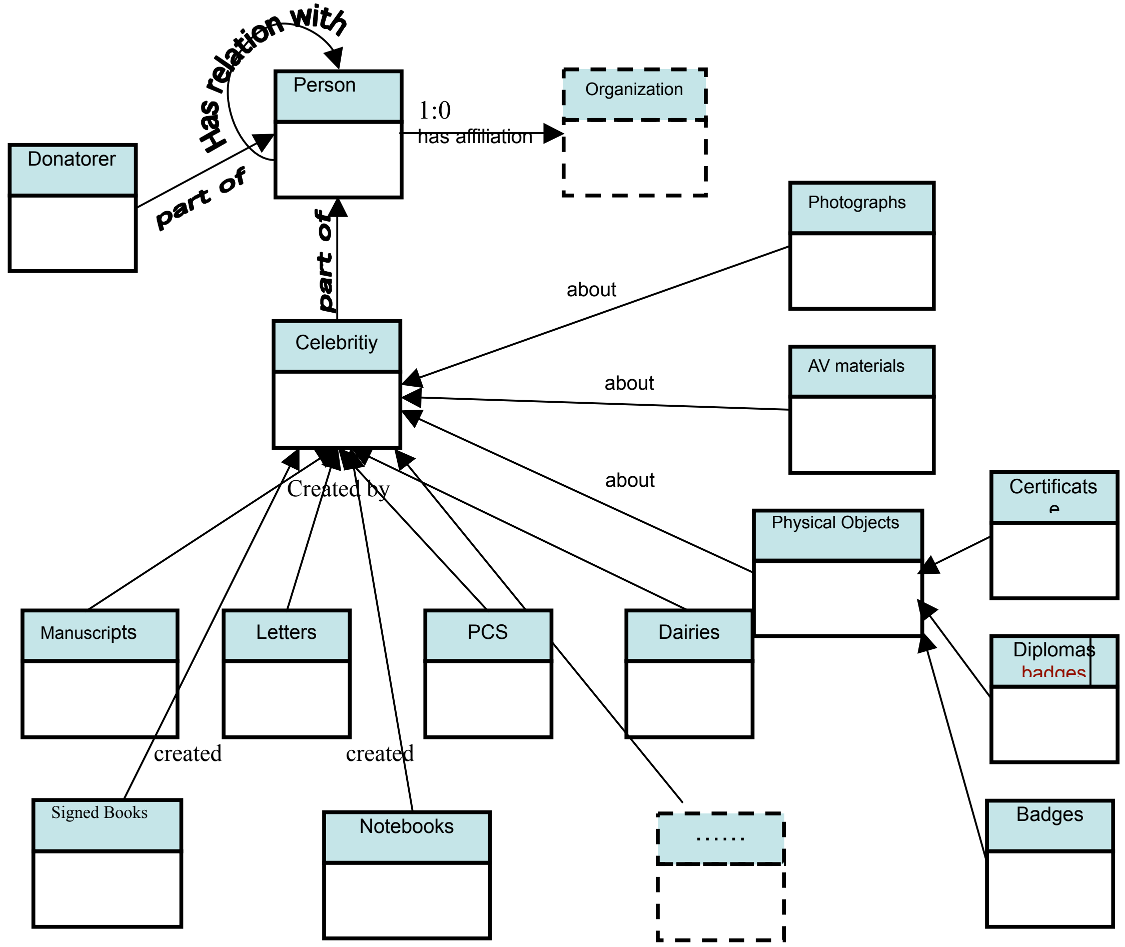
Another important distinctive characteristic of the CCCML collections is that each object in all kinds of resources is "about" a certain Celebrity. So related metadata elements which required by "related" resources such as the people, agent or institutions should be included.

In a word, the collections of CCCML are different from any traditional library or museum collections; the metadata schema for CCCML is therefore different from the metadata schema for digitized library or museum collections. To conform to general metadata standards, the metadata schemas have to compromise between accuracy/particularity and interoperability/generality And also, the metadata schemas for CCCML have to meet the individual needs by the requirements from both physical objects and digital content, such as the content management, long-term preservation, resource description, access control, semantic discovery, object circulation and so on, while at the same time, they have to be compatible with the whole digital library architecture of Shanghai Library, as well as with the national standards that is under establishment and international standards in this area. Besides, the processing procedure must be easy to control for staffs of the content professionals. The flexibility, interoperability and extensibility of the schemas are very important to the system.

[1] see http://www.malvine.org/

https://doi.org/10.23106/dcmi.952107854

Figure 1_ER model of the resources of CCCML



4. Methodology

Metadata schema is usually a joint effort by experts of computer specialists, content experts (librarians) and the users of the system. The conventional approach is to bring these people together to shape the detailed requirements for each type of the resources, then to turn an element set which consists of core elements and their qualifiers to a metadata schema. A schema developed in such a way often fails to maintain the consistency with the "future" standards, such as the national standards, which is still under development, and the evolving standards of Shanghai Library itself. It seemed to be an impossible mission to have compatibility or consistency with the evolving standards. However by obeying a few principles supporting the flexibility and extensibility of the system can help to do so. Principle 1: use mature models as the base to set up the system architecture. For example, OAIS[2] is an information system model that provides a comprehensive framework for systems especially dealing with the preservation issues; and FRBR[3] constitute a framework focusing on the relations among different forms in a lifecycle of a digital object. It is very useful to establish the ER model for a complicated system, which can be considered as an ontology model for the information system.

To summarize, the metadata schema of CCCML is a mixed Application Profile, using OAIS, FRBR, and DCMI's Abstract Model as methodology, adopting "Shanghai Library

[2] Reference Model for an OpenArchival Information System (OAIS)_Consultative Committee for Space Data Systems (CCSDS).URL_http://wwwclassic.ccsds.og/documents/pdf/CCSDS-650.0-B-1.pdf(access date_2004-2-1)

[3] Functional Requirements for Bibliographic Records (FRBR)_IFLA Study Group on the Functional Requirements for Bibliographic Records.URL_http://www.ifla.og/VII/s13/frbr/frbr.pdf (access date_2004-2-1)

https://doi.org/10.23106/dcmi.952107854

metadata schema" that is based on DC-Lib[4] application profile as its core metadata elements, and borrowing the elements from multiple metadata schemas and standards. The elements borrowed from other standards or schemas remained their original semantics without expansion or intersection, but it has some qualification or refinement to the specific vocabulary based on the requirements of the type of resources. The metadata application profile can be finally formalized partially by some standards of XML Schema such as METS[5] and MODS[6], or RDFs such as WSDL[7] in applications.

The procedure to generate a practical metadata schema can be illustrated as figure 2.

[4] Library Application Profile_Rebecca Guenther_URL_http://dublincore.org/documents/library-application-profile/(access date_2004-1-14)

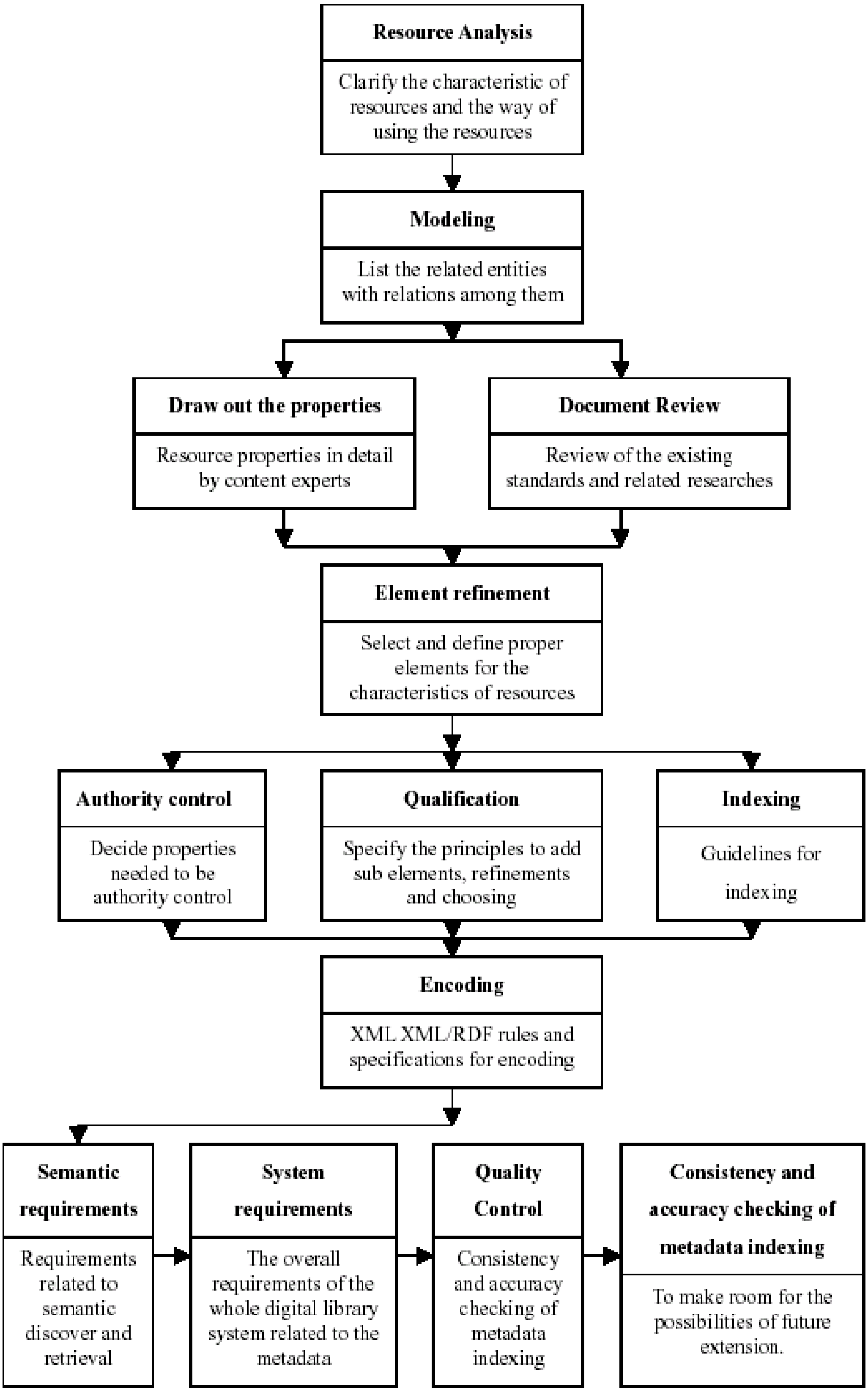
[5] Metadata Encoding and Transmission Standard (METS)_URL_http://www.loc.gov/standards/mets/(access date_2004-1-14)

[6] Metadata Object Description Schema (MODS)_URL_http://www.loc.gov/standards/mods/(access date_2004-2-1)

[7] Web Services Description Language (WSDL) 1.1_W3C Note_URL_http://www.w3.org/TR/wsdl(access date_2004-2-1)

https://doi.org/10.23106/dcmi.952107854

figure 2:The procedure to generate a practical metadata schema


Resource Analysis
Clarify the characteristic of resources and the way of using the resources
Modeling
List the related entities with relations among them
Draw out the properties
Resource properties in detail by content experts
Document Review
Review of the existing standards and related researches
Element refinement
Select and define proper elements for the characteristics of resources
Authority control
Decide properties needed to be authority control
Qualification
Specify the principles to add sub elements, refinements and choosing
Indexing
Guidelines for indexing
Encoding
XML XML/RDF rules and specifications for encoding
Semantic requirements
Requirements related to semantic discover and retrieval
System requirements
The overall requirements of the whole digital library system related to the metadata
Quality Control
Consistency and accuracy checking of metadata indexing
Consistency and accuracy checking of metadata indexing
To make room for the possibilities of future extension.

https://doi.org/10.23106/dcmi.952107854

5. The Semantic Architecture

5.1 The Purpose

Semantic architecture brings structure to the content of a digital library. The structure can expose some interfaces to outside world accessed by people as well as mediator agents. The design of semantic architecture is to give a practical approach under the consensus of semantic interoperability within and between communities.

We see the main purpose to establish a semantic architecture is to formalize the semantic description of digital resources, for the better serving of resource and service discovery, and exposing adequate interfaces for the integration of digital resources, and finally to achieve high level interoperability between digital libraries.

5.2 The Approach

The specification of Metadata Application Profile (MAP) provides the foundation of a semantic architecture for digital libraries. MAP is defined as a kind of metadata schema which consists of data elements drawn from one or more namespaces, combined together by implementers, and optimized for a particular local application [17]. It becomes a standard approach with methodologies and procedures to reuse metadata terms from various metadata standards authorities, share the semantics and structures all in once without the burden of setting up one's own metadata registry. One example of MAP is a CEN standard: CWA14855- "Dublin Core Application Profile guidelines", which is a declaration specifying which metadata terms to use and how these terms have been customized or adapted to a particular application. But it stopped in terminology level which can help to share a common data model underlying the applications but not information model which specifies complex relations among resources and properties during its life cycle.

The use of controlled vocabularies (thesauri), authority files and ontologies are practical means in the system level to achieve consistency and integrity within and between digital libraries. To get better flexibility and extensibility, especially in large institutions or enterprises with a number of various kinds of information resources and applications, the metadata registries which collect and maintain data dictionaries, metadata elements, schemas and vocabularies are the sources and repositories of formal semantics. They are the key mechanism to the semantic architecture, especially when the registries can provide web services for software agents on the request of digital library applications.

5.3 The Implementation

The semantic architecture for CCCML consists of schemas in data model level (which consists of the formal definition and restrictions of "core" elements, extended elements, metadata profiles, schema encoding rules) and information model level (which consists of relations between elements, ontologies, procedures and methodologies and Institutional registry for local qualified terms, schemas and namespaces), which serves for consistent description and discovery of semantics of the resources in CCCML. The architecture takes the form of a collection of schemas, tools and documentations which support semantics manipulation needs within the life cycle of the resources. The following paragraphs introduce the semantic architecture of CCCML system in a sequence of workflow:

1. Resource analysis and definition

The resources in the CCCML collections are defined from a practical point of view, from which the system can never anticipate what a set of properties of next object will be. We predefined twelve categories of resources with fixed metadata set and encoding schema in a form of Metadata Application Profile. But the system can accept multiple number and any kinds of MAP at the same time in the form of DTD, XML Schema or RDF Schema. The only necessity is the category of a resource should be defined explicitly with a set of properties (metadata elements from multiple namespaces with definitions), guidelines for cataloguing and encoding.

2. Metadata set definition (core and extended)

Shanghai Library has issued a specification with a "Core" set of metadata elements and encoding guidelines for the interoperation of all its digital library applications. The specification derived elements from DC-Lib application profile and consulted "the IFLA Guidance on the Structure, Content, and Application of Metadata Records for Digital Resources and Collections"[8]. As a digital library application of Shanghai Library, the CCCML system takes the

[8] See: http://www.ifla.org/VII/s13/guide/metaguide03.pdf

https://doi.org/10.23106/dcmi.952107854

"Core" as its mandatory set of elements. But this does not mean every element should be in use with the resource of CCCML. An element in the "Core" only becomes mandatory when it is needed.

At the same time each type of resources in CCCML 'borrows' some elements from other metadata standards like MODS, VRACore etc., and proposes its own elements, as its the domain specific MAP. So a local metadata registry should be established to maintain terms in a local namespace for the proposed elements as well as for those terms from other metadata standards without namespaces.

It is not recommended to invent elements or terms for any of the resources. But the content owners and users of CCCML want to discover the properties of the resources exhaustively. So we developed a rigorous procedure for approving the proposed terms.

3. Encoding and mapping Rules

The Schema Suite is a stand alone utility to manipulate (open, load, input, parse, edit, save, delete, convert, output etc.) metadata schemas and generate web interfaces for metadata cataloguing as well as help to generate the query interface. It is designed to support DTD, XML Schema and RDF Schema according to the rules of encoding from time to time. All empty schemas (without instances) can be kept and managed with the tool.

Basically the tool is fed with an original schema of the "Core" set. But it supports aliases for core elements so that it can be userfriendly to the domain expert for inputting and retrieving of the resources. It can support to accept records with ISO2709 format and transform it to any form of a MAP according to a mapping table.

4. Guidelines and Best Practices

Due to the limitation on the capabilities of different formalization language like XML Schema or RDF Schema, not all of the restrictions and constrains can be expressed and encoded with them. Some of the functions have to accomplish during implementation.

So the metadata element set and its encoding are not enough to carry the semantics of an information model. It must assist with restrictions, constrains, rules, guidelines etc. That's why the semantic architecture has documentation for people readable instead of machine readable. All these documents would better be kept and maintained in a mechanism of registry system so as to provide open access by people or agents. What's more it can be extended to construct web service to provide semantic support services (discussed below).

5. Metadata registry, Ontologies and Authority Files

Registries are essential to the scalability of a digital library, for it provides a mechanism in the distributed environment to get the semantic architecture reusable, sharable, integrity and consistent. Local registry is a kind of "have-to" facility for institutions and enterprises as the scale of application becomes bigger and bigger and eventually get out of control. Registry can be considered as data dictionary for local systems. But the metadata registry should synchronize with open registries distributed on the internet. And it's better open to serve as a member of the metadata registry cluster.

6. The Metadata Schema

Metadata schemas for CCCML are the main output of the semantic architecture, which set up the specifications for metadata indexing.

To design a specific metadata schema is different from the making of metadata standards, which should include a series of documents along with the procedure of metadata implementation. We include six documents as necessary guidelines for the CCCML application:

Table 2_The metadata schema documents

Resource analysis report
The metadata element set and the definition (usually called MAP: Metadata Application Profile)
Cataloguing/indexing Rules
Encoding schema
Authority control
The requirements of system

The metadata element set defined with a format conformed to ISO11179. Twelve properties have been defined. Each element is listed below:

- Chinese identifier
- English identifier
- Namespace
- Edition
- Registry
- Language
- Obligation
- Definition
- Description
- Data type
- Maximum Occurrence
- Encoding scheme and

As mentioned above, the core elements contained 17 of 18 elements from Metadata Schema Specification of Shanghai Library which is based on DC-Lib. The whole element set consists of

https://doi.org/10.23106/dcmi.952107854

core elements and extended elements. But not all the core elements are obliged to be existed in the schemas for all kinds of resources. The schema for CCCML borrowed five elements from CDWA[9] (they are: Materials and Techniques, Cataloging History, Facture, Related Textual Reference, Inscriptions/ Marks), one element from REACH[10]_Place of Origin/Discovery_ , and introduced several elements as needed. The schemas allows alias in accordance with resource type. The element set for CCCML is the union of all element sets for all kinds of resources in CCCML. In general, if a new type of resource is adopted, it's better to choose the elements from the existing element set so as to remain stability of the whole system.

5. Problems

It is difficult to implement such a hybrid comprehensive metadata schema for CCCML. No ready solution can be used. One problem is that the same element has alias when it is used to describe different type of resource. The user friendly customization provides the domain specific interface to content experts while remaining basic interoperability. For example, the CREATOR of resource "LETTERS" has a name of "DONOR". However, in the system, the different names of the same element must be consistent with the "core" elements so as to ensure the coherency of the metadata schema. Another problem is, an element for describing different resources should have different refinements or different encoding schemes. The system should be able to deal with this problem at runtime.

Conclusion

The metadata schemas for CCCML are implementations of Metadata Specifications of Shanghai Library. But as discussed above, in the metadata schemas for CCCML, there are no administrative metadata, preservation metadata and technical metadata, which will be fully implemented during the realizing of the CCCML digital library.

Reference_

1. MALVINE:Manuscripts and Letters via Intergrated Networks in Europe_URL_http://www.malvine.org/
2. Reference Model for an Open Archival Information System (OAIS)_Consultative Committee for Space Data Systems (CCSDS)_URL_http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf (access date:2004-2-1)
3. Chinese Metadata Schema(proposal)_National Library_ 2001-6
4. Functional Requirements for Bibliographic Records (FRBR)_IFLA Study Group on the Functional Requirements for Bibliographic Records.URL_http://www.ifla.org/VII/s13/frbr/frbr.pdf(access date:2004-2-1)
5. Dublin Core Abstract Model_Andy Powell_URL_http://dublincore.org/documents/abstract-model/(access date:2004-1-14)
6. RDF Vocabulary Description Language 1.0: RDF Schema_W3C Proposed Recommendation_URL_http://www.w3.org/TR/rdf-schema/(access date:2004-2-1)
7. Library Application Profile_Rebecca Guenther_URL_http://dublincore.org/documents/library-application-profile/(access date:2004-1-14)
8. Guidance on the Structure, Content, and Application of Metadata Records for Digital Resources and Collections_URL_http://www.ifla.org/VII/s13/guide/metaguide03.pdf(access date:2004-1-14)
9. Metadata Encoding and Transmission Standard (METS)_URL_http://www.loc.gov/standards/mets/(access date:2004-1-14)
10. Metadata Object Description Schema (MODS)_URL_http://www.loc.gov/standards/mods/(access date:2004-2-1)
11. Web Services Description Language

[9] Categories for the Description ofWorks of Art(CDWA)_URL_http://www.getty.edu/research/conducting_research/standards/cdwa/(access date_2004-2-1)

[10] RLG REACH Element Set for Shared Description of Museum Objects _URL_http://www.rlg.org/reach.elements.html(access date_2004-2-1)

https://doi.org/10.23106/dcmi.952107854

(WSDL) 1.1_W3C Note_URL_http://www.w3.org/TR/wsdl(access date:2004-2-1)

12. The Handbook for the Research Work of Specific Digital Object Description Metadata Standard_revised_.Specific Digital Object Description Metadata Standard Working Group_Interior material_2003-7
13. Categories for the Description ofWorks of Art(CDWA)_URL_http://www.getty.edu/research/conducting_research/standards/cdwa/ (access date:2004-2-1)
14. RLG REACH Element Set for Shared Description of Museum Objects _URL_http://www.rlg.org/reach.elements.html(access date:2004-2-1)
15. Xiaolin Zhang.Metadata Research and Application_Beijing_Beijing Library Publishing Company,2002-05
16. Xuehua Chen, Zhaozhen Chen, Guanghua Chen. XML/Metadata Management System for Digital Library. Taipei: Wenhua Library Management Information Company, Minggo90[2001]
17. Thomas Baker, Makx Dekkers, Rachel Heery, Manjula Patel, and Gauri Salokhe, "What Terms Does Your Metadata Use? Application Profiles as Machine-Understandable Narratives". Journal of Digital Information, Volume 2 Issue 2 (November 2001)

https://doi.org/10.23106/dcmi.952107854