Challenges in Digital Libraries - Key Issues Learned from Metadata-Centric Projects at Tsukuba

Shigeo Sugimoto
Research Center for Knowledge Communities
Graduate School of Library, Information and Media Studies
University of Tsukuba, Tsukuba, Ibaraki 305-8550, Japan
email: sugimoto@slis.tsukuba.ac.jp

ABSTRACT

The digital library has been widely recognized as an important infrastructure for information and knowledge sharing in the networked information society. This paper aims to discuss some key issues in the research and development of digital libraries based on experiences in the metadata-centric digital library research projects in which the author has been involved. The projects discussed in this paper involved the development of a few subject gateways and the metadata schema registry in collaboration with the Dublin Core Metadata Initiative (DCMI). From the subject gateway projects, he learned that community oriented metadata schemas, especially metadata vocabularies, have a very important role and that an appropriate technology to develop and maintain the metadata schemas is required. From the metadata schema registry project with DCMI, the author has learned the importance of a conceptual framework of metadata which helps us better understand semantics and structure of metadata schemas, and augment their interoperability. This paper first gives an overview of digital libraries briefly. It then discusses some key issues in digital libraries based on the experiences learned in these digital library projects. The key issues for digital libraries are interoperability, model and theory, preservation, accessibility, metadata and so on. This paper also explains some basic concepts of metadata schema and briefly describes some of the projects and lessons learned. Then, the paper discusses some basic concepts of metadata and the metadata-centric projects.

INTRODUCTION

The World Wide Web has explosively expanded over the world during these ten years and we use not only PCs but also PDAs and mobile phones to access the Internet, e.g. to send/receive emails and to access the Web. Our network infrastructure has also greatly progressed during these ten years – broadband connection is now widely available and wireless connection to the Internet is expanding. Thus, our information environment has drastically progressed and is changing day by day. The very rapid growth of information and communication technologies (ICT) will continue to evolve and enhance our networked information society.

The digital library is an essential service to help users find and access information resources in the networked information society. Many research and development projects on digital libraries have been carried out by the ICT research communities and by memory organization communities since the early 90's. The Digital Library Initiative (DLI) cooperatively organized by the National Science Foundation (NSF) and other governmental agencies of the United States was the most influential funding program for the research communities in computer and information sciences to develop new information technologies and new concepts of digital libraries [1]. The National Science Digital Library program started on projects to develop educational digital libraries [2]. In Europe, the digital library was also recognized as an important topic in the 5th and 6th Framework of EU. For example, "DELOS Network of Excellence on Digital Libraries" in Europe has been supporting advanced technology researches for digital libraries and related activities [3]. A broad range of research topics are included in the digital library research agenda as we can see in the research and development projects of digital libraries. Among the research topics, several crucial issues for the development of digital libraries have been identified through the various projects. NSF and DELOS organized working groups to discuss key issues for the development of digital libraries [4][5]. The topics discussed by the working groups are metadata, interoperability, multilingual information access, information discovery, intellectual property, digital preservation, actors in digital libraries, etc.

Memory organizations such as libraries and museums have greatly contributed to develop new services of digital resources in the networked information environment, e.g., digitization of rare/historical resources,

development of scholarly resource repositories, development of hybrid library environments, among others. For example, American Memory of the Library of Congress is a very early adopter of the Web to make their historic collections accessible through the Internet [6]. Publishing, especially the scholarly publishing environment, has greatly changed during these ten years. Many commercial publishers and academic societies are publishing their journals electronically via the Web and they are providing new services such as virtual journals and cross-referencing. There are many repositories of scholarly resources such as pre-prints, technical reports and dissertations. The Open Archives Initiative, for instance, is a collaborative effort by repositories to connect the repositories and add value to them [7].

The author has been involved in research projects in digital libraries at the University of Library and Information Science (ULIS) and University of Tsukuba[1]. The topics of the projects include subject gateways, a metadata schema registry in collaboration with the Dublin Core Metadata Initiative (DCMI), metadata models, digital archives, and so forth. The key and common component among these projects is metadata. One critical issue learned from these projects was the gap between requirements of different communities, i.e. difference of requirements for the global audience and regional/domain-specific audience. The Internet is the global information infrastructure and many different communities use the Internet to publish and share information resources. This means that we need to satisfy both requirements for global interoperability and those for a specific community, which is very challenging to do at the same time.

This paper intended to discuss crucial issues in digital libraries learned mainly through the projects in which the author was involved. This paper discusses the issues mainly from a technological point of view. In this paper, the author shows some lessons learned from these activities. Before doing that however, the author first provides an overview of the research and development activities on digital libraries. Secondly, he discusses several key topics for digital libraries mainly from a technological point of view. Then, this paper shows the projects at Tsukuba and discusses some lessons learned from the projects.

DIGITAL LIBRARIES – AN OVERVIEW

Some Key Topics and Activities in digital Library Research

Research and development activities in digital libraries are roughly classified into two categories - technology-centric projects and service-centric projects. Typical examples of the former are the projects funded by DLI. On the other hand, many projects of the latter category, which are primarily designed for development of digital collections and digital library service, have been carried out mainly by libraries, e.g., LoC's American Memory.

One of the most notable points of technology-oriented projects is its cross-disciplinary feature. Many large-scale projects involved participants from different communities; i.e., researchers in computers, telecommunications, library and information sciences/technologies, humanities and social sciences, practitioners at memory organizations such as libraries and museums, and stakeholders of resources for digital libraries. For example, an archeological digital library would require knowledge of the subject domain of archeology to collect and organize resources and information technologies to digitize and present the resources properly.

National libraries and major research libraries have been leading the library community. They have invested a large amount of their resources for the development of digital collections and their technologies. In the scholarly publishing community, electronic journals have been widely accepted and many scholarly journals are published both in printed and electronic forms. On the other hand, huge volumes of scholarly resources such as pre-prints, technical reports and dissertations published on the Web have been collected and made available online by repositories. Grassroots communities have also contributed lots of resources on the Web. For example, Aozora Bunko[8], which is a voluntary and not-for-profit activity in Japan, has built a collection of electronic texts of 4000 titles of Japanese novels.

Several key issues for the development of digital libraries have been identified through these research and development activities; for example, interoperability among digital libraries, information resource discovery across language and cultural borders, accessibility and adaptability in accordance with characteristics of users and their environments, long term preservation of electronic resources, intellectual property issues, security issues, and so forth. Various digital library technologies are required to solve

[1] ULIS was merged with University of Tsukuba in 2002 and became the Graduate School of Library, Information and Media Studies.

these issues. Metadata technology is an important and common component to solve these key issues.

Digital Library Development at and for Libraries

Our information seeking behaviors have been changed by Internet search engines and Internet resource directories, e.g.,Google and Yahoo! In the last ten years, broadband networks and wireless connection to the Internet have greatly progressed in Japan. As of 2004, 62% of network connection at home is via broadband, whereas it was only 6.8% in 2000. There are about 80 million Internet users in Japan as of 2004. Among these 80 million users, 42 million people use both PCs and mobile phones to access the Internet, while 21 million and 15 million people use only either PCs or mobile phones to access the Internet [9]. This large population of Internet users from mobile phones is the distinctive feature of the Japanese information access environment.

This change has heavily affected library services. For example, it is common for libraries to provide their services via the Internet, e.g. OPAC, digital collections and reference services. In Japan, more than 1000 public libraries provide OPAC services via the Web and about 200 libraries provide their homepages for Web access from mobile phones [10].

Libraries have made large efforts to build digital library services via the Internet. The list below shows typical digital library functions/services provided by libraries.

(1) To collect and provide resources published in digital forms, e.g., electronic journals and databases.
(2) To collect resources and organize them as digital collections, e.g., digital collections of rare and historical materials, or collections of digitized and born digital resources.
(3) To provide information about information resources and help users find and access resources, e.g., subject gateways.
(4) To provide reference services via the Internet.
(5) To connect digital libraries to provide larger and value-added collection of resources.
(6) To preserve and archive digital contents, e.g., Web archives.
(7) To provide a user-friendly environment to use all kinds of resources, e.g., personalization of digital library service and hybrid library.

KEY ISSUES FOR DIGITAL LIBRARIES

A digital library is a large-scale integrated system which provides various types of services to various types of users. There are several crucial issues for digital libraries, e.g., intellectual property, security, metadata, digital preservation, resource accessibility by users with/without disabilities, interoperability among digital libraries, model of digital libraries, etc. The following paragraphs discuss from a technological point of view some of the issues.

(1) Interoperability

Interoperability among digital libraries is crucial in order to enhance accessibility and usability of networked resources across borders irregardless of user characteristics, e.g., user's mother language, age and disabilities. Interoperability is crucial not only to find and access information resources but also to browse and interact with the resources. Preservation, which is in other words interoperability over time, is also widely recognized as crucial for digital libraries.

(2) Model and theory

A formal framework to figure out digital library structure is crucial to understand the features of digital libraries, i.e., system structures, functional and service requirements, administrative and management structures, etc. For example, the 5S model by Fox defines a layered model of digital libraries [11]. The model provides us with a framework to conceptually understand the organization of a digital library and to identify its components for purposes of interoperability among digital libraries.

(3) Digital Preservation:

Preservation of digital resources, especially born digital resources, is challenging technologically and socially. Digital preservation, including Web archiving, is currently one of the major research issues in digital libraries, especially for deposit libraries and archives. We need to preserve not only the resources but also the technologies and environments required to use the resources in order to perfectly preserve them. However, this is a very difficult requirement to achieve. Aside from this fundamental issue, digital preservation has several challenging issues which are as follows:

(a) Selection of resources for preservation

Resource selection for preservation is primarily a cost-effectiveness and policy issue. In addition, it includes a technological issue to correctly identify an instance of a resource to be collected for preservation. For example, in the case of preserving an XML text which is associated with a set of

style-sheets, a policy to determine how the primary XML resource should be preserved, e.g., whether the look-and-feel of the resource should be preserved, and if technologies to identify what information resources should be preserved in addition to the primary XML resource.

(b) Consistency and integrity of resources as a collection and as a single resource

Consistency and integrity of a single resource have to be maintained in an archive. A policy to determine the consistency and integrity of a preserved resource is primarily required since a resource could be composed of more than one resource component connected by hyperlinks and the boundary of a resource is not always clear. Then, we need a technology to identify a "single" resource. Consistency and integrity management of a collection of resources is crucial as well.

(c) Coverage and consistency of automated collection of resources

Ordinary Web crawlers collect resources via the Internet asynchronously with the creation and update of the resources. In addition, there are Web pages whose access is controlled and not open to general public, i.e. hidden Web. This means that cooperation between Web-site managers and Web crawler managers is required in order to augment the coverage of automated collection and consistency of the collected resources.

(d) Metadata for preservation

Resources have to be preserved with appropriate set of metadata for preservation and discovery in and across the archives. Some authorities have defined metadata schemas for preservation based on the Open Archival Information System [12]. Detailed metadata is desirable but cost-effectiveness of metadata description has to be examined for implementation [13].

(4) Services for global and regional/domain–specific communities – Globalization vs. Localization, Universality vs. Domain-Specificity

The Internet is a global information infrastructure. Digital libraries provide their services over the Internet, which means that digital libraries have general requirements to provide their services for the global community. On the other hand, each digital library has its own regional/domain-specific requirements for its target audience, e.g. collection building, classification schemes, and so forth. Thus, digital libraries need to satisfy requirements both for global and regional/domain-specific communities.

(5) Accessibility and adaptability of resources in accordance with users and user environments

Accessibility has been broadly recognized as an important aspect to provide digital library services over the Internet. Digital libraries need to lower the barriers for any user to access resources. Technologies to find resources in accordance with user characteristics and user environments and to adapt the resources in accordance with the user are required.

(6) Metadata

Metadata, which is defined as "data about data", is widely recognized as one of the key issues for digital libraries. Metadata is an important component to realize library services in the networked information environment since user services, such as search and access to resources, heavily rely on metadata of resources. Several metadata standards and related technologies have been developed during this decade, e.g., Dublin Core, Metadata Object Description Standard (MODS), Metadata Encoding and Transmission Standard (METS), Learning Object Metadata (LOM), Resource Description Framework (RDF), and so on. Traditionally, metadata schemas have been defined by application or by community. On the other hand, the World Wide Web provides a framework of metadata on the Web, i.e. RDF/XML and related standards, as an infrastructure to share metadata among different communities. The author considers that the following issues have to be taken into account to design metadata schemas for digital libraries.

(a) Interoperability and reuse of metadata schemas

Developers of digital library services would define metadata schemas in accordance with their own requirements as determined by the type of applications, type of resources, type of users and user environments, and so forth. However, they need to pay attention to the interoperability of metadata with other services, as well as reuse of existing metadata schemas.

(b) Development and maintenance of metadata schemas

A metadata schema for a community is designed in accordance with the requirements of the community. This means that the community would require community-specific components in the metadata schema. On the other hand, community-specific components have to be maintained by the community, which could be a burden for the community.

In order to solve these issues, the author believes that a good conceptual framework of metadata schemas is required. The Dublin Core community has contributed fundamental concepts in this aspect. In this paper, the author shows a simple conceptual model of metadata schema in the next section.

BASIC CONCEPTS OF METADATA SCHEMA

Before describing metadata-centric research, this section briefly explains basic concepts of metadata schema and some key concepts of metadata schemas developed by the Dublin Core community which are crucial to understand the basic features of metadata schemas used in the Internet.

Metadata Schema

Metadata is defined as "data about data" or "structured data about data". In the library community, metadata is a central component of library services because the data created by libraries to manage library holdings and to provide library services is mostly metadata. A metadata schema defines a framework of representation of a metadata. In general, a metadata schema includes semantic definition of terms used in the schema, structural constraints and data structure definitions, and bindings to physical description syntax such as XML.

A metadata schema consists of the following components:

(1) a set of terms defined to express properties of a resource, e.g., *Title*, *Creator*, *alternative* and so on,
(2) a set of terms which expresses types of property values and/or which are used as a property value, e.g. *ISO-8601*, *DCMI Type Vocabulary*, *LCSH*, and *DDC*,
(3) a set of rules which defines structural constraints and syntactic features neutral to any implementation specific description scheme, e.g. mandatory levels, repeatability/cardinality, order, and so on, and
(4) a set of binding rules to a specific description language.

Cataloging rules, in general, include guidelines for catalogers to extract values from resources to create catalogs in addition to the components listed above. The definition of metadata schema in this paper does not include the guidelines.

For example, Simple Dublin Core has the following metadata schema constructs in terms of the items listed above:

(1) It has a set of 15 elements.
(2) No specific vocabulary is given but some widely used vocabularies and standards are recommended.
(3) It has a weak structural constraint that is "every element is optional and repeatable", and
(4) The concrete syntax, i.e., representation in a specific language, is not included in Simple Dublin Core. The bindings to HTML, XML and RDF are given in separate documents.

Basic Concepts of Metadata Schema in Dublin Core

This section describes some basic concepts of the model and the Dublin Core process. The DCMI Abstract Model gives the precise underlying data model for the DCMI metadata [14].

(1) Warwick Framework and Application Profile

Since the Internet is a very diversified environment, it is useless to assume that a single metadata element set will meet the needs of all domains and purposes. It is also impractical to develop metadata sets application by application: the result would be expensive and chaotic, and interoperability would be non-existent. On the other hand, it is desirable for application developers to use established metadata schemas and adopt them in accordance with local requirements. The Warwick Framework, a conceptual model that resulted from the 2nd Dublin Core Workshop in 1996, gave an early expression to the notion of metadata as modular components that may come from more than one metadata schema [15]. In this model, a metadata instance is expressed as a container which contains one or more packages, each of which is expressed in a given metadata schema. The Resource Description Framework (RDF) provided a practical realization of many of the ideas of the Warwick Framework.

Application Profiles, which provide a framework to adopt one or more element sets in accordance with an application, could also be considered a realization of the Warwick Framework [16]. Dublin Core Metadata defines the vocabulary of metadata, i.e., terms and their meanings, but in general does not specify the encoding or syntactic characteristics. An exception is the feature included in Simple DC that is "Any of the 15 elements is optional and repeatable." Local applications, however, may have domain specific requirements appropriate to a given domain or application:

- Title, Creator and Description might be mandated but others are optional,
- Use only Title, Creator, Description, Date and Language elements,
- Use the 15 elements of Simple DC and some elements from other metadata sets such as the IEEE Learning Object Metadata (IEEE LOM), and so forth.

These requirements can be defined independently of the vocabulary definitions. Description of this application-specific syntactic feature is called an application profile. Any application can have its own

application profile, which specifies a set of metadata vocabulary terms used in the application as well as syntactic or structural features of the particular application. Figure 1 shows a model of application profiles. The vocabulary terms could be borrowed from one or more source schemas. More importantly, the application profile could be used to define a mapping between the application's scheme to a global scheme(s), which is crucial for interoperability.

(2) Dumb-down Principle

The Dumb-Down principle gives a guideline for qualification. The Dumb-Down principle suggests that a value of a qualified element has to be consistent as a value of the element without any qualification. For example, assume the following qualified values:

(1) (Element Refinement) Date Accepted: "2004-10-12",
(2) (Encoding Scheme) Language: "en" encoded in RFC 1766, and
(3) (Value Structure) Creator: {name: "Sugimoto, Shigeo", affiliation: "University of Tsukuba", contact: "sugimoto@slis.tsukuba.ac.jp"}

Then, assuming that the qualifications in the above examples, *Accepted*, *RFC 1766* and the component names of the value structure (i.e., *name*, *affiliation* and *contact*) are removed. The values of example 1 and 2, "2004-10-12" and "en" are still consistent with their elements after the removal. However, the value of example 3 {"Sugimoto, Shigeo", "University of Tsukuba", "sugimoto@slis.tsukuba.ac.jp"} causes problems since the second and third values are not valid values of *Creator*.

Dumbing-down is a crucial function for metadata interoperability in the global community since local communities can extend their schemas in accordance with their requirements, and at the same time they can also keep their metadata interoperable with other metadata communities.

(3) Evolution and Maintenance of Metadata Vocabularies and Schema Registry

Any existing metadata standard needs its process model to keep the standard updated in accordance with the requirements given to the standard. The following paragraphs show the maintenance process model of metadata terms by DCMI.

To remain relevant in a rapidly evolving Web environment, Dublin Core must be able to grow and evolve in response to user needs. DCMI has therefore instituted a Usage Board and a process model for reviewing proposals for expanding or clarifying the standard. Proposed elements and element refinements that conform to Dublin Core principles are taken into the standard with the status of *conforming*. To some proposed terms of proven usefulness for resource discovery across domains the Board may assign the status of *recommended*. Proposals for encoding schemes are reviewed for accuracy and given the status of *registered*. Once approved, each new term is assigned a Uniform Resource Identifier using one of the official namespace URIs maintained by DCMI. A "namespace policy" defines limits within which the metadata terms maintained by DCMI can evolve or change over time. According to this policy, editorial changes or updates are allowed, but changes of semantics (meaning) are not; new semantics require the creation of new element.

DCMI metadata terms are stored in the DCMI metadata schema registry and its cooperating registries.

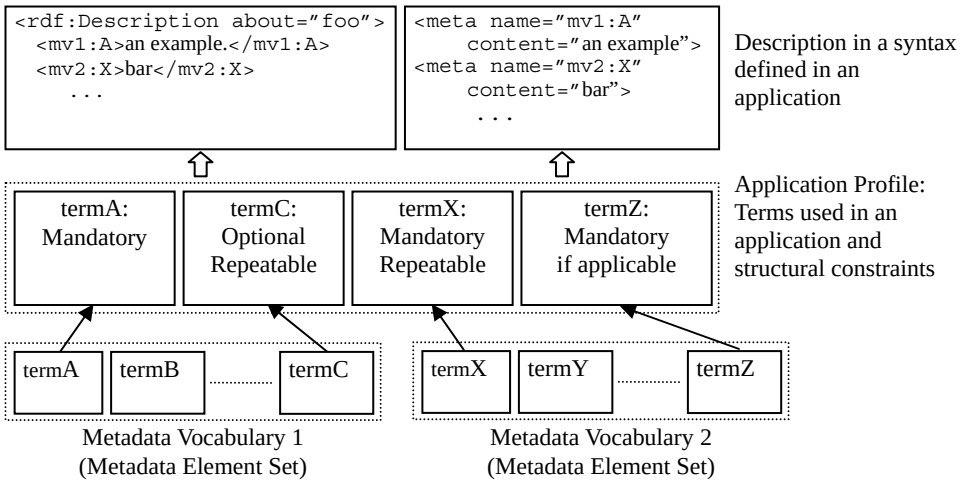

Figure 1. Concept of Application Profile

The terms are made accessible via the Internet and maintained in the registries. Authoritative reference descriptions of the metadata terms in English are translated into non-English languages for adoption of local communities. By the nature of Dublin Core, this translation of the vocabularies has been and will be done by grassroots volunteers. In addition, a local community can define its own metadata terms, which may or may not be approved as c*onforming*. Metadata vocabulary maintenance has to be performed in two aspects; one is the authoritative description directly maintained by the Usage Board, and the other is translations in non-English languages. The authoritative description is stable but, on the other hand, a translated description is rather unstable unless it is translated by a local authority.

SUBJECT GATEWAY PROJECTS AT TSUKUBA – BUILDING COMMUNITY-ORIENTED SUBJECT VOCABULARIES

This section describes metadata issues learned from three metadata-centric projects at Tsukuba. Each of the projects is designed for a domain-specific community (ULIS-DL), for children and in multiple languages (IPL-Asia), and for regional resources (Digital Okayama Dai-Hyakka), respectively. The following sections describe metadata issues in these projects, especially on the subject classification vocabularies in these projects. The projects are reported in [17][18][19].

ULIS-DL

The principal purpose of ULIS-DL is to build a subject gateway to resources useful for libraries and LIS institutions. We have collected the resources published mainly by libraries and LIS institutions in Japan, and created metadat for the resources. The metadata element set called ULIS Core, is defined based on the 15 Simple Dublin Core elements with a few ULIS-DL specific elements.

We have developed a small subject vocabulary in order to build a directory-style navigational interface for ULIS-DL which shows subject terms sorted in a hierarchical order and a list of resources associated to every subject term. A preliminary evaluation of the Subject element values showed that there are more than 15,000 distinct text strings in the raw metadata as of summer 2003, which includes typographical errors, inappropriate use of upper/lower case letters and so on. We also found that a set of subject terms assigned to a page in a Web site significantly overlaps to that of other pages in the same site and that the divergence of the number of metadata records per site is significantly large.

After having the raw metadata normalized, we created a candidate core subject vocabulary by extracting terms that appear two or more times in the set of normalized metadata records. We found that approximately 90% of the total records is covered by a set of about 1000 terms each of which appear five times or more. We classified the terms of this set into eight categories which are (1) Web terms, e.g., links, (2) Library terms, e.g. OPAC, (3) Organization and facility information, e.g. floor guide and access, (4) Type of libraries, e.g. university library and public library, (5) Organization names and service names, (6) Place names, (7) General subject terms, and (8) Reference tools, e.g. dictionaries, thesauri. Then, we classified terms in these categories into sub-categories up to the third level to constitute a hierarchical structure of subject terms. We assigned a proper subject term to each node of the tree and encoded the classification vocabulary in OWL.

A Subject Gateway in Multiple Languages - Internet Public Library Asia

Internet Public Library Asia (IPL-Asia) started in the year 2000 at ULIS, which was initially planned partly as a collaborative activity with the University of Michigan. In order to develop IPL-Asia, we first formulated some criteria for Internet resource selection and a metadata schema. Based on these criteria, we collected resources written in Chinese, Japanese and Korean (CJK) languages. Metadata was assigned for the resources in CJK and also in English based on the metadata schema which are chosen from those of the Dublin Core Metadata Element Set (DCMES) and Learning Object Metadata (LOM). Each metadata record was collaboratively created by a group of catalogers.

As part of this project, we adopted UDC as a subject vocabulary. However, we learned that a subject vocabulary for IPL-Asia need not be comprehensive but it has to be defined in accordance with the application domain and the audience. That is because the domain of the resources is narrow and the subjects are community-specific, the subject terms of those well-established vocabularies are difficult for children to understand, and appropriate terms should be chosen to express the subjects in accordance with the age levels of the children. A single concept should be expressed in different forms in accordance with the age of the audience. For example, in the case of Japanese audience, a subject term for children of first to third grade would need to be written only in Hiragana with a limited set of Kanji characters. On the

Table 1. Distribution of regional subject vocabulary terms in the NDC term space

The upper rows of KV and PV show the numbers of terms mapped to corresponding NDC class and the lower rows show their ratios. A single KV or PV term is mapped to one or more NDC terms. "Number of NDC Terms" means the number of distinct NDC terms in each major category. (NDC major categories: 000=Generalities, 100=Philosophy, 200=History, 300=Social Sciences, 400=Natural Sciences, 500=Technology, 600=Industry, 700=The Arts, 800=Language, 900=Literature)

	NDC categories	000	100	200	300	400	500	600	700	800	900	total
KV	**#Terms**	**17**	**8**	**8**	**196**	**58**	**54**	**28**	**62**	**6**	**6**	**443**
293	**Ratio (%)**	**3.8**	**1.81**	**1.81**	**44.2**	**13**	**12**	**6.3**	**14**	**1.35**	**1.35**	100
terms	**#NDC terms**	**7**	**7**	**3**	**44**	**27**	**27**	**20**	**26**	**4**	**4**	169
PV	**#terms**	**15**	**2**	**12**	**171**	**30**	**56**	**44**	**17**	**1**	**1**	**349**
287	**Ratio (%)**	**4.3**	**0.6**	**3.4**	**49**	**8.6**	**16**	**13**	**4.9**	**0.3**	**0.3**	100
terms	**#NDC terms**	**4**	**2**	**5**	**34**	**11**	**18**	**25**	**15**	**1**	**1**	116

other hand, children of junior high school and high school ages would prefer the term expressed using an ordinary set of Kanji characters. Thus, a single concept could have multiple human readable labels. Vocabulary description languages such as RDF Schema, XML TopicMaps, and OWL have a function to assign multiple human readable labels to a single concept.

Another lesson we learned was the cost to create metadata. Metadata was primarily created in a single language and then translated into other languages which was time consuming. This project is reported in detail at the DC-2003 conference in Seattle[20].

Digital Okayama Dai-Hyakka (DODH) and its Subject Vocabularies

Digital Okayama Dai-Hyakka (DODH) is a regional portal by the Okayama Prefectural Library in collaboration with other public sectors in the prefecture. DODH provides a Z39.50-based OPAC across public libraries in the prefecture of Okayama, a reference database, and Okayama Regional Information Network (ORIN) which provides a gateway to regional resources. ORIN uses a metadata schema based on Simple Dublin Core. ORIN uses three subject vocabularies - a classification scheme for general resources, a classification scheme for the resources published by the prefectural government and Nippon Decimal Classification (NDC). The first scheme, which is called Okayama Kids Vocabulary (KV), is designed primarily for the general public and children. The second scheme, which is called Okayama Prefecture Vocabulary (PV) in this paper, is designed for resources including Web pages created by the prefectural government. Both of these subject vocabularies are designed to be sufficiently simple since the subject terms will be used by the general public and children, and because metadata will be produced by non-professional catalogers. On the other hand, NDC is used by librarians. The authors contributed in the design of KV based on the experiences in IPL-Asia.

NDC terms used in this system include the major 1000 categories. Three mapping tables for all pairs of these three categories were created. All of the mappings between these three subject vocabularies were created by OPL. Mapping between terms is not 1:1, i.e. a single term in a vocabulary is mapped to one or more terms in another vocabulary. Table 1 shows the distribution of KV and PV terms against NDC terms. This table shows that social sciences (300) category is heavily mapped both from KV and PV. Figure 2 outlines the distributions. This figure shows KV is oriented to natural science and arts but PV is oriented to industries in addition to social sciences. This seems natural because KV is oriented towards children resources and educational resources and, on the other hand, PV is oriented to resources published by the prefectural government of Okayama.

The subject terms of KV has four presentation labels chosen in accordance with user ages, i.e. first to third graders (junior level of elementary school), fourth to six graders (senior level of elementary school), seventh to ninth graders (junior high school level), and eighth or higher graders (high school to general public). Presentation labels are determined in accordance with age of the audience as we did in IPL-Asia – easy terms expressed in only syllabic characters (Hiragana and Katakana) for the youngest group and ordinary terms for high school children and higher.

Summary of the Issues Learned from the Metadata-Centric Projects

The paragraphs below summarize a few key issues learned through the projects:

(1) Type and Granularity of Resources: A general goal of subject gateways is to help users find useful resources. A subject gateway developer collects resources which are useful in its subject domain and creates metadata for the resources. This process looks similar to cataloging of conventional library materials but the fundamental difference between them is the diversity of type and granularity of the resources – metadata can be created for a whole site, a single page or even a single file. Thus, metadata schema for a subject gateway should be designed in accordance not only with the domain of the subject gateway but also with the type and granularity of the resources.

(2) Controlled Vocabularies: Vocabularies for classification and subject description of networked information resources are an important component to build digital library services. There are vocabularies such as DDC and LCSH which are broadly adopted for conventional resources but the projects mentioned above required reasonably small vocabularies tailored to their resources and users.

(3) Metadata Schema Sharing: Interoperability is a very important aspect for digital library services. Sharing information about metadata schemas is an important step to achieve interoperability. Sharing metadata schema information is also important to encourage people to adopt and/or customize existing schemas in order to build a new schema. Therefore, technologies to promote sharing of metadata schema information are a crucial issue.

SIMPLE CONCEPTUAL MODEL OF METADATA SCHEMAS AND SCHEMA REGISTRY

A Layered Model of Metadata Schema

As described in a previous section, metadata schema includes semantic and syntactic components. These components can be organized into layers as follows:

Layer 1 - Semantics Definition Layer (or ontology layer): Definition of terms used in the schema. In other words, definition of metadata vocabulary, i.e. metadata element set. In general, two types of metadata terms are included in the metadata vocabulary – property vocabulary and value vocabulary[21]. A property vocabulary or, in other words, element vocabulary, is a set of property terms, for example, elements and element refinement qualifiers of DCMES. A value vocabulary is a set of value terms, for example, encoding schemes of DCMES. Definition of each term should primarily include a primary name and its meaning. Thus, a vocabulary definition gives the semantic basis of a metadata schema.

Layer 2 - Structural Constraints Definition Layer (abstract syntax layer): Definition of syntactic features which does not depend on any particular implementation scheme. A set of terms used in the schema and structural constraints applied to each term should be included in a definition. Application profiles are given in this layer. The structural constraints would include composition, ordering, mandatory levels, repeatability and cardinality, and specification of controlled vocabularies used in a metadata element. In other words, this layer defines application profiles in implementation neutral syntax.

Layer 3 - Implementation Dependent Syntax Definition Layer (concrete syntax layer): Definition of syntax of metadata in an implementation; for example, metadata description syntax in HTML, XML, RDF or in a specific database management system such as Oracle and MySQL.

Figure 3 illustrates a layered model which is based on a single element set, i.e. DCMI Metadata Terms. Simple Dublin Core specifies “use of the 15 elements of Dublin Core where every element is optional and repeatable.” As shown in Figure 1, an application profile in layer 2 can be defined based on multiple metadata element sets.

An application schema developer would provide guidelines for creating metadata in addition to their schema. The guidelines can be documented in layers 2 and/or 3 in accordance with the implementation

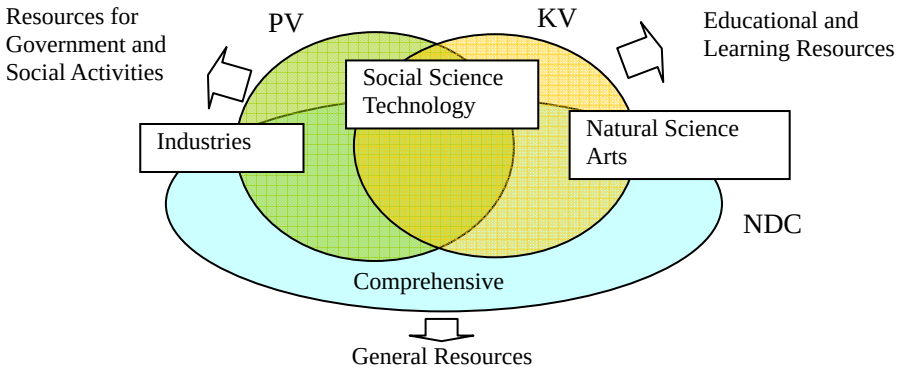

Figure 2. Outline of Coverage of Vocabularies

specificity; for example, the DCMI Library Application Profile includes some general guidelines in implementation neutral level, which should be associated to layer 2.

A metadata term defined in layer 1 can be defined in an ontology specification language such as RDF Schema and OWL. Structural constraints in the layer 2 can be defined in a syntax description scheme such as DTD, RELAX NG and XML Schema.

A Simple Requirements Analysis for Metadata Interoperability based on the Layered Model

The layered model helps us better understand requirements to realize retrieval functions across different metadata schemas. The following paragraphs show very simple requirements analysis cases for retrieval across metadata repositories.

Case 1: Repositories A and B have the same metadata schema in all layers. Metadata instances of both repositories are interoperable as they are.

Case 2: Metadata schemas of A and B are the same in layers 1 and 2. This case needs common implementation syntax. Conversion from the original physical syntax to the common syntax should be straightforward.

Case 3: Metadata schemas of A and B use the same vocabularies defined in layer 1 but syntactic features in the higher layers are different. This case needs extraction of commonly used metadata terms and definition of a set of metadata terms as an interoperability set.

Case 4: Metadata Schema of A and B partly share vocabularies in layer 1. This case needs extraction of a common set of terms and definition of an interoperability set. For the extraction, dumb-down function could be applied.

Case 5: Metadata schema A and B have no common vocabulary. This case needs definition of crosswalks between A's and B's vocabularies for creating a common set of metadata terms and a common syntax of metadata instances.

In the practical environment, more detailed requirements analysis would be required; for example, metadata extraction guidelines and processes should be examined to check if a metadata element is used in a narrower meaning without defining a refined element and so on.

In the requirements analysis above, metadata vocabulary gives the basis for metadata interoperability. Formal definition scheme of metadata vocabulary should be used to create descriptions of metadata term definitions that have to be both machine and human understandable. RDF Schema has been used by the DCMI metadata schema registry as a formal vocabulary description scheme. In RDF Schema description, every metadata term is given a unique identifier which works as its primary name. A term definition could include one or more secondary names and related information as well. The primary name – typically a URI – is defined to uniquely identify the term. On the other hand, since secondary names are given as a human-friendly label, the secondary names could be translated into other languages from English. The primary names are used in the formal specification of metadata schemas to identify metadata terms and other constructs and to define relationships between them.

Since his simple analysis is based on the underlying model of Dublin Core and RDF and description scheme of structural constraints is not explicitly included in the model this simple analysis model does not include the analysis of structural constraints in a metadata schema. Structural constraints are classified into the following types:

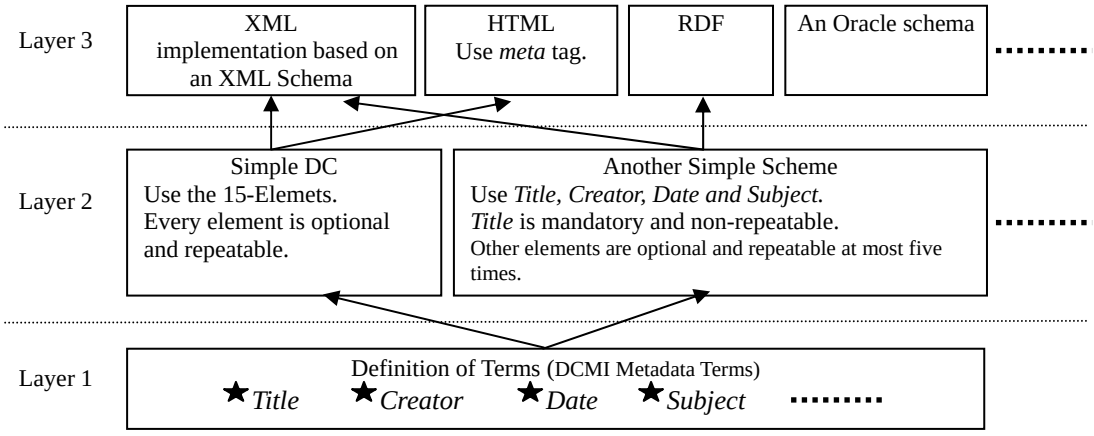

Figure 3. Layered Model of Metadata Schema

(1) A composite value composed of named sub-elements. For example, a person name composed of a first name, a given name, an affiliation, and a contact address.
(2) A composite value composed of ordered/unordered sequence of component values. For example, an ordered list of component values whose minimum and maximum lengths are 5 and 10, respectively.
(3) Mandatory levels, e.g., optional, recommended, mandatory if applicable, mandatory.
(4) A set of value types of an element adopted in the application, which should be a proper set of the value types of the element defined in layer 1.
(5) Ordering constraints, i.e., descending or ascending order of values or significance of values, e.g. list of authors.

In general, mapping of metadata structures between different schemas needs structural transformation on a case-by-case basis. It is possible to define a generic function for the transformation, e.g., a function to structurally dumb-down a composite value into a simple value which conforms to the schemas, and a function to extract elements which are common among the schemas being cross-used. On the other hand, some information could be lost during this transformation.

Metadata Schema Registry

DCMI has a metadata schema registry which stores reference descriptions of the DCMI terms. Each DCMI term is encoded in RDF Schema. Every definition of a term includes a unique identifier, a label(s), a description(s), a comment(s), relationships to other terms and some more information. The labels, descriptions and comments are primarily expressed in English and then translated into non-English languages. As of June 2005, the DCMI registry provides the reference description in 25 languages.

The DCMI registry is developed primarily to store the DCMI terms but it is extensible to any metadata vocabularies. The registry at Tsukuba stores metadata vocabulary of IPL-Asia and some other metadata vocabularies. We have experimentally applied the Tsukuba registry to develop software tools which use metadata term definitions and cooperate with the registry [22].

DISCUSSION AND CONCLUDING REMARKS

From the projects and activities in digital libraries, the author has learned key technological issues mentioned in the third section. In addition, the topics described in the following paragraphs show some of the organizational and human resource issues. In general, IPR and security issues are crucial for digital libraries but excluded in this paper because of the author's capability and area of interests.

(1) Collaborative development and maintenance of digital libraries

Since geographical distance has no significant meaning on the Internet, collaboration among libraries to build and augment digital library services is crucial to share resources and enhance usability of digital libraries. Libraries can decrease costs to collect the information about the resources in the virtual space and they can add values in accordance with their own requirements and environments.

(2) Adaptation to new information and communication technologies and environments

Digital libraries need to catch up with new technologies in order to develop resources and user environments and to adapt them for their services. For example, digital libraries would need to provide interface not only for PCs but also mobile phones because huge number of people are using mobile phones as the primary Internet access device. Ubiquitous information technology, which is a hot topic for ICT research and development in Japan, would affect conventional and digital libraries.

In Japan and in other developed countries, the information infrastructure has been constructed and is progressing every day. This means that digital libraries need to evolve day by day. On the other hand, this progress of the information infrastructure means that the diversity of users is expanding. There are lots of challenging goals left for future development of digital libraries. In addition, libraries should not forget the digital divide issue, and those users who live on the other side of the digital divide.

Acknowledgements

The author would like to express his sincere thanks to Prof. Koichi Tabata for his support and contributions to these projects. He would like to thank Dr. Stuart Weibel of OCLC, Dr. Thomas Baker of Goettingen State and University Library, and Mr. Mitsuyoshi Moriyama of Okayama Prefectural Library for their cooperation with the projects. Last but not least, he would like to express his gratitude to all of his colleagues and students who contributed to the studies described in this paper.

References

[1] Digital Library Initiative Phase 2, http://www.dli2.nsf.gov/
[2] National Science Digital Library, http://nsdl.org/

[3] DELOS Network of Excellence on Digital Libraries, http://www.delos.info/ (Sixth Framework Programme), http://delos-noe.iei.pi.cnr.it/ (Fifth Framework Programme)
[4] Joint NSF-EU Working Groups on Future Directions of Digital Library Research, http://www.dli2.nsf.gov/internationalprojects/workgroups.html, 1999
[5] International Forum: DELOS/NSF Joint Working Groups, http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/joint-wgs.html, 2003
[6] American Memory, http://memory.loc.gov/ammem/
[7] Open Archives Initiative, http://www.openarchives.org/
[8] Aozora Bunko, http://www.aozora.gr.jp/ (in Japanese)
[9] Ministry of Internal Affairs and Communication, “White Paper Information and Communications in Japan (Japanese Version)”, Chap.1, 2005, accessible from http://www.johotsusintokei.soumu.go.jp/whitepaper/ja/cover/index.htm
[10] Japan Library Association, “Web Site Services by Public Libraries”, http://www.jla.or.jp/link/public2.html (in Japanese)
[11] Gonçalves, M.A., Fox, E.A., Watson, L.T., Kipp, N.A., “Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries”, ACM Transactions on Information Systems (TOIS), Vol.22, Issue 2, pp.270-312, 2004
[12] Consultative Committee for Space Data Systems, “Reference Model of an Open Archival Information System (OAIS)”, 2002, http://www.ccsds.org/docu/dscgi/ds.py/Get/File-143/650x0b1.pdf
[13] Calanag, M.L., Tabata, K., Sugimoto, S., “Linking Preservation Metadata and Collection management Policies”, Collection Building, Vol.23, No. 2, pp56-64, 2004
[14] Powell A., et al., “DCMI Abstract Model”, http://www.ukoln.ac.uk/metadata/dcmi/abstract-model/
[15] Lagoze, C., “The Warwick Framework – A Container Architecture for Diverse Sets of Metadata”, D-Lib Magazine, July/August 1996, http://www.dlib.org/dlib/july96/lagoze/07lagoze.html
[16] “Dublin Core Application Profile Guidelines”, CEN Workshop Agreement (CWA) 14855, 2003
[17] Sugimoto, S., Baker, T., Weibel, S., "Dublin Core: Process and Principles", Proceedings of ICADL 2002, Singapore, 2002 (LNCS 2555, pp.25-35, Springer)
[18] Sugimoto, S., “Metadata Schemas, Models and Tools - Metadata-Centered Projects at Tsukuba and Lessons Learned for Interoperability”, Proceedings of ICDL'04, India, 2004, pp.690-699
[19] Sugimoto, S. et al., “Developing Community-Oriented Metadata Vocabularies: Some Case Studies”, Proceedings of DLKC'04, 2004, pp.128-135
[20] Lee, WS., et al., “A Subject gateway in Multiple Languages: a Prototype Development and Lessons Learned”, Proceedings of DC-2003, pp.59-66
[21] Baker, T., et al., “Principles of Metadata Registries”, http://delos-noe.iei.pi.cnr.it/activities/standardizationforum/Registries.pdf
[22] Nagamori, M., Sugimoto, S., “A Metadata Schema Framework for Functional Extension of the Metadata Schema Registry”, Proceedings of DC-2004, pp.3-11, 2004