

A Subject gateway in Multiple Languages: a Prototype Development and Lessons Learned

Wonsook Lee, Shigeo Sugimoto, Mitsuharu Nagamori, Tetsuo Sakaguchi, Koichi Tabata
University of Library and Information Science (ULIS)
Institute of Library and Information Science, University of Tsukuba, Japan
{wonsook, sugimoto, nagamori, saka, tabata}@ulis.ac.jp

Abstract

The Internet has a lot of rich information resources available in different languages aside from English, the major language of use. Those resources are, however, not easy to access because of language barriers. This paper proposes a collaborative model to develop a multilingual subject gateway, which is called the Internet Public Library Asia (IPL-Asia). It describes how IPL-Asia was conceived and developed, and some lessons learned from the experience.

IPL-Asia is a subject gateway which collects high quality Internet resources expressed in Chinese, Japanese, Korean languages, and provides information about those resources in those three languages and English. This paper shows the metadata schema and an implementation of IPL-Asia. The schema is defined based on Dublin Core and IEEE LOM with some extensions for multilingual description. Criteria for selecting quality resources on the Internet were also defined. A metadata record is collaboratively described in multiple languages by catalogers whose mother languages are different. A metadata record is encoded in XML. Metadata records are stored and displayed using XML technologies, e.g., XSLT and XML database. This system provides user interfaces in multiple languages.

Keywords: Information Gateway, Subject Gateway, metadata, Collaborative Metadata Development, Multilingual Information Access

1. Introduction

The Internet has been widely spreading over the world since the mid-1990's, and it has become an essential part of our information and knowledge infrastructure. Internet users are widely diverse – ranging from children to senior people, from scholars and professionals to the general public. Internet resources are also diverse not only in their formats and their types of contents but also in their languages.

There are widely accepted services such as Google and Yahoo! to find resources on the Internet. However, the diversity of Internet users is one major reason for the difficulties of searching high quality resources on the Internet in accordance with user requirements. From this viewpoint, subject gateways, or information gateways or portals, have been recognized as crucial services on the

Internet for effective distribution and access to high-quality information resources.

Multilingual information access is recognized as one of the key issues for digital libraries [1]. The multilingual issue is obviously an important but challenging issue for subject gateways because there are many languages used for publishing resources on the Web. In general, information resources are published in a single local language or at most two, i.e., a local language and English, and in some cases in a few local languages. Subject gateways are mostly provided in a local language, and there are very few gateways which provide their services in more than one language.

It is not easy for Internet users to find and access those resources across language borders. Description of a quality resource is useful for a user even if the resource is expressed in a foreign language. Since the language barrier is much higher for children than adult users to find and access good Internet resources expressed in foreign languages, description of quality resources in a local language for children is crucial.

Having considered the above issues, we started to construct a multilingual subject gateway we called the 'Internet Public Library Asia (IPL-Asia)', which is partly a collaborative project with the Internet Public Library [2]. IPL-Asia is focused on resource collections published in Chinese, Japanese and Korean (CJK) languages. We collect high quality information resources published in CJK and provide information about those resources in those languages and in English. We have defined a metadata schema for IPL-Asia and experimentally developed an IPL-Asia service. This paper gives an overview of IPL-Asia and lessons learned from this project.

The core part of IPL-Asia is a set of metadata expressed in the four languages (i.e. CJK + English). A metadata record which includes description in more than one language is primarily done by a group of human catalogers whose mother languages are different. Metadata creation done manually is quite expensive; however, since quality of metadata description is crucial for subject gateways, and we feel that current state-of-the-art technology of automatic metadata generation and machine translation are not yet mature enough to be used by subject gateways for metadata creation. In this project, we have also explored designing a user interface with children in mind.

In Section 2, basic issues for subject gateways are mentioned. Section 3 shows the outline of IPL-Asia,

including the metadata schema and a set of software tools designed to support the development of IPL-Asia. Section 4 shows some lessons learned from this study and future work.

2. Subject Gateways

The primary service of subject gateways is to provide description of high-quality information resources. Generally speaking, there are two types of subject gateways. One is, as the name suggests, focused on a certain subject domain, and the other is one which covers multiple subject domains or a broad range of domains. Examples of the former type are AHDS [3] for arts and humanities, BIOME[4] for health and life sciences, SOSIG[5] for social science, Biz/ed[6] for business and economics, GEM[7] for educational resources and so forth. Examples of the latter type are RDN[8], Renardus[9], WWW Virtual Library[10], and IPL.

2.1. A Subject Gateway in Multiple Languages

Although the Internet is a global information network, Internet resources are primarily published in a local language in non-English speaking regions. Generally speaking, we can say that resources published in local languages are much richer than those published in English in the non-English speaking regions such as in Asia. In those regions, ordinary users primarily access resources published in their local languages especially if translations of English works are available in their languages before they turn to English resources, which are the largely the original sources of these translations.

Since the Internet is a global network, any resource published on the Internet is “globally accessible and sharable”. However, this means that resources are accessible and sharable only over geographic distance but not across language barriers. There is no easy answer to solve the language issue. Multilinguality is one of the key issues for subject gateways, but it has been mostly neglected because of its difficulty.

Subject gateways can provide information about resources regardless of the languages in which the resources are written. Resource description in multiple languages is a central part of a multilingual subject gateway. Quality of resource description is crucial for subject gateways to enhance their usability and reliability. In this respect, the descriptive component of a metadata instance, e.g. Description element in terms of Dublin Core, has a very important role for human users.

2.2. Metadata in Multiple Languages

Technologies to assist information access across language barriers are crucial to solve the multilinguality issue in subject gateways, e.g. machine translation and multilingual information retrieval. High-quality description in natural languages and appropriately controlled classification over multiple languages are

required to develop a multilingual subject gateway. However, state-of-the-art technologies for natural language processing and metadata creation have not caught fast enough to produce high-quality metadata description for a multilingual subject gateway. In this experimental study, we have developed a collaborative work model for a group of human catalogers to create metadata records in multiple languages. Based on the model, software tools to support the development and management processes were developed [11].

The model was proposed based on our experience in developing a multilingual collection of Japanese Old folktales [12][13]. The collection includes several Japanese folktales translated into 10 languages. Those tales were translated by volunteers collaborating over the Internet. In this task, the volunteers were native speakers of the target language of translation in order to ensure translation accuracy. We employed the same model for IPL-Asia. This quality requirement applies to multilingual metadata as well. Resource descriptions in natural language should ideally be done by a native speaker or by a specialist in the language for the sake of description quality.

2.3. Requirements Analysis of Multilingual Metadata for IPL-Asia

Metadata for a multilingual subject gateway needs to be looked at from various aspects: (1) metadata schema, (2) metadata creation (i.e., metadata description), (3) organization and maintenance of metadata, and (4) user interfaces of the subject gateway. The metadata schema includes both language-dependent elements and language-neutral elements. For example, Title and Description elements are language dependent, and Date and Format elements are language neutral, in general.

Since IPL-Asia is primarily designed for the general public and children, its metadata schema should include elements from the viewpoint of education and learning in addition to general resource description elements. A metadata record is to be created and maintained by a group of catalogers whose mother languages are different. A language-neutral software environment should be used to organize metadata records as a database. In this respect, Unicode and XML technologies are useful. User interfaces have to be provided in multiple languages, although logically there is only a single metadata database. User interface design for adult users and children is another important language issue to consider. What can be done, for example, is that subject terms in a directory could be presented according to age.

3. Internet Public Library Asia

3.1 Overview

Internet Public Library Asia started in the year 2000 at the University of Library and Information Science (ULIS) in Tsukuba City, Japan as a collaborative project

with the Internet Public Library (IPL), which is based at 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 CJK. Metadata was assigned for the resources in CJK and also in English based on the metadata schema. Each metadata record was collaboratively created by a group of catalogers. The user interface of IPL-Asia was also prepared in multiple languages.

The IPL-Asia system was developed using XML technologies. A metadata record is encoded and stored in XML and converted into HTML by XSLT for presentation in browsers.

3.2 Metadata Schema

Metadata elements and qualifiers of the IPL-Asia metadata schema are chosen from those of the Dublin Core Metadata Element Set (DCMES)[14] and IEEE LOM [15]. As shown in Table 1, the metadata schema has 17 elements, some of which have refined elements expressed by element refinement qualifiers. In other words, the schema is an application profile defined based on DCMES and IEEE-LOM. Some of the elements are language-neutral, but descriptive elements such as Title

and Description contain language-dependent values. Each language-dependent element contains a value encoded in more than one language in parallel.

We adopted the metametadata element from IEEE LOM in addition to DCMES elements. The metametadata element is used primarily for maintenance purposes but it has an important role in collaborative metadata creation by a group of catalogers. Some of the elements are given qualifiers defined for IPL-Asia. For example, the Description element has “long” and “short” qualifiers. These qualifiers are adopted in order to make translation task easier. This is because short descriptions are often hard to translate and would sometimes lose useful information. On the other hand, a short description makes it easy for users to recognize useful resources. Another example is the Audience element which has the Agerange qualifier to specify target audiences of the resource.

Metadata description guidelines for catalogers were created based on the resource selection criteria and the general guidelines provided by DCMI. We have also adopted the Nippon Cataloging Rules (NCR) as bases for the guidelines with some extensions, e.g. notation rules of foreign names. The metadata schema and the guideline documents were prepared in Japanese. In this experimental study, all metadata records were created by a group of graduate and senior students of ULIS whose common language was Japanese.

Table 1 Metadata Application Profile of IPL-Asia

Notes:

DR: Description Requirement (Mandatory or Optional)

ES: Original Element Set names (DC: Dublin Core, LOM: IEEE LOM, IA: IPL-Asia).

Elements and qualifiers		DR	ES	Description
Title		M	DC	A name given to the resource.
	Main-Title	M	IA	Main Title.
	Sub-Title	O	IA	Any non-main title, e.g. alternative Title.
Creator		O	DC	An entity primarily responsible for making the content of the resource.
Publisher		M	DC	An entity responsible for making the resource available.
	Current	M	IA	Name of an entity responsible for making the resource available.
	Old	O	IA	Names other than current name.
Identifier		M	DC	An unambiguous reference to the resource within a given context.
Description		M	DC	An account of the content of the resource.
	Long	M	IA	An account of the contents of the resource; typically longer than 7 lines.
	Short	M	IA	An account of the contents of the resource; typically about 3 lines.
Type		M	DC	The nature or genre of the content of the resource.
Format		M	DC	The physical or digital manifestation of the resource.
Language		M	DC	A language of the intellectual contents of the resource.
Date		M	DC	A date of an event in the life cycle of the resource (a date of an event in the life cycle of the resource or adoption in IPL-Asia).
Metametadata		M	LOM	The description of metadata record itself.
	Entry	M	LOM	Creation date of the metadata record.
	Language	M	LOM	A language of the contents of the metadata, e.g., Chinese, Japanese, Korean and English.

	Contributors	M	IA	A person or a group of persons who created and are responsible for the metadata record.
Audience		M	DC	A class of entity for whom the resource is intended or useful (typical users of the resource, i.e. general, teen or youth in IPL-Asia).
	Agerange	O	IA	Target age range of the resource; especially required for textbooks.
Subject		M	DC	The topic of the contents of the resource.
	UDC Code	O	IA	UDC encoded subject.
Coverage		O	DC	The extent or scope of the content of the resource.
	Temporal	O	DC	Temporal coverage of the content of the resource.
	Spatial	O	DC	Spatial coverage of the content of the resource.
Relation		O	DC	Relation to other resources; with qualifiers specifying the type of relationship.
	IsVersionOf	O	DC	The described resource is a version, edition, or adaptation of the referenced resource; changes in version imply substantive changes in content rather than differences in format.
	HasVersion	O	DC	The described resource has a version, edition, or adaptation, namely, the referenced resource.
	IsReplacedBy	O	DC	The described resource is supplanted, displaced, or superseded by the referenced resource.

	Replaces	O	DC	The described resource supplants, displaces, or supersedes the referenced resource.
	IsRequiredBy	O	DC	The described resource is required by the referenced resource, either physically or logically.
	Requires	O	DC	The described resource requires the referenced resource to support its function, delivery, or coherence of content.
	IsPartOf	O	DC	The described resource is a physical or logical part of the referenced resource.
	HasPart	O	DC	The described resource includes the referenced resource either physically or logically.
	IsReferencedBy	O	DC	The described resource is referenced, cited, or otherwise pointed to by the referenced resource.
	References	O	DC	The described resource references, cites, or otherwise points to the referenced resource.
	IsFormatOf	O	DC	The described resource is the same intellectual content of the referenced resource, but presented in another format.
	HasFormat	O	DC	The described resource pre-existed the referenced resource, which is essentially the same intellectual content presented in another format.
Source		O	DC	A reference to a resource from which the present resource is derived.
Rights		O	DC	Information about rights held in and over the resource.
	Cost	O	IA	Description about cost to use the resource.
	Restrict	O	IA	Description about restrictions and conditions to use the resource.
	Copyrights	O	IA	Description about copyrights.
Contributor		O	DC	An entity responsible for making contributions to the content of the resource.
	Editor	O	IA	Editor of the resource.
	Translator	O	IA	Translator of the resource.
	Illustrator	O	IA	Illustrator of the resource.
	Designer	O	IA	Web-designer of the resource.

3.3. Multilingual Metadata and its Description Process Model

In development of the prototype, we have collected metadata for two kinds of resources: Internet resources for children expressed in CJK, and collections of historical materials digitized by a few public libraries in Japan. Figures 1 and 2 show a multilingual view and a

monolingual view of a metadata record, respectively. Figure 1 shows a multilingual view of a metadata record, which contains descriptions in CJK and English. Figure 2 shows a monolingual view (in Japanese) of the metadata record.

In our metadata creation model, a metadata record is collaboratively created by a group of catalogers whose mother languages are different. A metadata of a resource

is first assigned by a cataloger whose mother language is the language of the resource. Next, the original description is translated into other languages by his/her collaborating catalogers. The catalogers communicate with each other via a common language(s) among them. In this case, Japanese is the primary common language. The catalogers do not have to understand the other languages other than his/her mother language and the common language for collaboration.

Element	Korean	Japanese	Chinese	English
TITLE				
MAIN-TITLE	Tour2 Korea.com	Tour2 Korea.com	Tour2 Korea.com	Tour2 Korea.com
SUB-TITLE	한국관광공사	韓国観光公社	韩国观光公社	Korean National Tourism Organization
PUBLISHER				
CURRENT	한국관광공사	韓国観光公社	韩国观光公社	Korean National Tourism Organization
IDENTIFIER	http://www.kato.or.kr/Korea/index.html	http://www.kato.or.kr/Korea/index.html	http://www.kato.or.kr/Korea/index.html	http://www.kato.or.kr/Korea/index.html
DESCRIPTION	한국 관광 공사의 홈페이지로, 구성은 크게 원형 서비스, 멀티 미디어 서비스, 여행 코너로 되어져 있다. 원형 서비스 한국 관광 공사가 보유하고 있는 관광 정보 데이터 베이스를 대략 1개 반 정도 이상으로 관광 정보 이용 하는 편의를 대상을 관광 정보 및 시설 정보를 모아와 정보를 제공하기 시킬 것이다. 멀티 미디어 서비스에서는 사진 공간으로의 수장품과 공사의 정보 비디오, 오디오, 홍보 영상, 관광 발족책 등을 이용할 수 있다. 여행 코너에서는 관광지, 토산품, 음식, 스포츠 시설, 레저, 음악지, 호수, 교통, 관광 상품, 조수 연차치 별로 검색을 할 수 있다. 그리고 지역별 검색과, 행사안내, 관광 정보 등도 제공하고 있다. 그리고, 추천할 만한 여행지, 지역별 관광 코스를 한국을 여행하기 전 꼭 필요한 정보와 만만 해준다. 한국에 들어와 여행도 잘라, 일본어, 독일어, 에스파냐어, 프랑스어, 중국어와 한국어가 사용되어져 있다.	한국관광公社のホームページで、構成は大まかに原形サービス、マルチメディアサービス、旅行コーナー、公社の新聞、お楽しみサービスから成っている。原形サービスは韓国観光公社が所有している観光情報データベースで、公社の観光情報を必要とする人々を対象として会員を募集して情報を提供するサービスである。マルチメディアサービスは、写真公開展の受付と公社の観光ビデオ・CM、観光ビデオ等を利用できる。旅行コーナーは、観光地、文化施設、スポーツ施設、レストラン、飲食店、ショッピング、交通、観光商品、重要観光地、非常災害、その他、その他、その他、旅行案内、観光地間の観光情報も提供している。韓国観光のページ以外に、英語、日本語、ドイツ語、エスパニッシュ語、中国語のページも用意されていて、その内容は異なっている。	韩国国家旅行公社的主页是公司由经营状况信息服务、多媒体服务、旅游角、公司新闻、热情接待等部分组成的。经营状况信息服务是韩国国家旅行公社为韩国需要公司旅行信息的会员们提供所提供的公司的旅游信息数据库。多媒体服务是韩国国家旅行公社为韩国需要旅游信息的人们提供照片、广告、旅游日历、旅游指南等对观光地、文化设施、运动设施、休闲地、饮食店、购物、观光纪念品、重要观光地等分别进行检索，并且提供旅游地地图的检索。旅行角中，观光地图等。除此之外，还有推荐观光地、地下停车场设施等。在出行韩国之前，此页面提供事先须知。除了提供韩国国内的旅行之外，还提供有中国、日语、德语、西班牙语、英语的主页，但内容有所不同。	The homepage of the Korean National Tourism Organization is broadly composed of the following categories: tourism information service, multimedia service, travel guide, newsletter and recommendations. The tourism information service, mainly used by its corporate members, distributes tourism information from the organization's database. Multimedia service users can access a gallery of prize-winning photos from a tourism photo festival, the organization's promotional video and tourism advertisements, as well as a calendar of events. Users of the travel guide can search for information under the major subregions of travel including shopping facilities, venues for leisure and sporting events, restaurants, shipping, transportation, and souvenirs. Users can also search for event guides and tourist maps based on geographic region. Also displayed are major tourist sights and subway routes; these pages are definitely worth visiting at least once before visiting Korea. In addition to Korean, Chinese, English, Japanese,

Figure 1. IPL-Asia Metadata Example

	C:\Program Files\Microsoft Internet Explorer\IEXPLORE.EXE		7/26/2012 12:09 PM
	http://www.knto.or.kr/korean/event.html		
要素	日本語		
タイトル	Tour2 Korea.com		
現在の形態にした者	韓国観光公社		
識別子	韓国観光公社		
内容	http://www.knto.or.kr/Korean/index.html 韓国観光公社のホームページで、源泉サービス、マルチメディアサービス、旅行案内、公社の新聞、お勧めのコーナーのカテゴリがある。そしてこの以外の様々なカテゴリで韓国の旅行情報を検索することができる。各情報欄も得られる。韓国語のページ以外にも英語、日本語、ドイツ語、スペイン語、中国語のページが用意されていて、その内容は異なっている。 韓国観光公社のホームページで、構成は大まかに源泉情報サービス、マルチメディアサービス、旅行案内、公社の新聞、お勧めのコーナーから成っている。源泉サービスは韓国観光公社が所有している観光情報のデータベースを、公社の観光情報を作成する法人を対象として会員を集めて情報を提供するサービスである。マルチメディアサービスでは、写真公募展の実件と公社の広報ビデオ(CM、観光カレンダー等)が利用できる。旅行案内では、観光文化施設、スポーツ施設、レジャー、飲食店、ショッピング、交通、観光商品、重要施設等別に検索でき、その他には地域毎の検索、旅行案内、観光地図の情報なども提供している。それからお勧めの観光地、地下鉄観光コースなど、韓国を旅行する前に一度チェックしてみる価値があるページである。韓国語のページ以外にも英語、日本語、ドイツ語、スペイン語、中国語のページが用意されていて、その内容は異なっている。		
カテゴリー			

Figure 2. Metadata record displayed in a single language - Japanese

In this model, the Description element is a very important information source for secondary catalogers who translate metadata descriptions from the description done by the primary cataloger. Precisely for this purpose, a “Long Description” is made part of the schema in order to include rich information in a metadata record even though it is a labor- and time-intensive task.

The following shows the process of multilingual metadata development:

- (1) Resource Collection: collecting high quality Internet resources expressed in CJK.
- (2) Metadata Creation: writing metadata for the resources in the four languages.
- (3) Metadata Organization: organizing metadata records as a database for service.

- (4) Metadata Service and Maintenance: providing the subject gateway service and maintaining the metadata database.

The first two steps are collaborative work by a group of people. The first step is a collaboration between a resource specialist who finds quality resources and a cataloger(s) who prepares metadata description for the resources. The second step is a collaboration among catalogers. The third and fourth steps are similar as those in an implementation process of regular monolingual subject gateways. Since the second step of translating metadata is time-consuming, we have found that metadata records that have not completed their translations be used for the service as well. Translations of the metadata record could be added later to the database when they are made.

3.4. Multilingual User Interface and Metadata Database

IPL-Asia provides user interfaces in multiple languages, i.e. Chinese, Japanese, Korean and English. All pages except the entry page are written in one of the four languages. The entry page is written in four languages and provides entry points to pages written in one of the four languages. All pages are encoded in XML (Figure 3) and converted into HTML using XSLT.

```
<title>
<main-title>
<langstring contents="original" xml:lang="ko">
사춘기의 특성</langstring>
<langstring contents="translation" xml:lang="ja">
思春期の特性</langstring>
<langstring contents="translation" xml:lang="zh"> 青春期的特性</langstring>
<langstring contents="translation" xml:lang="en"> The
characteristic of adolescence</langstring>
</main-title>
</title>
```

Figure 3. Part of an XML-encoded metadata record

In storing these metadata records, we do not recommend any specific database tool. Any off-the-shelf database tool which accepts Unicode and XML texts can be used. IPL-Asia used PostgreSQL as its database tool.

3.5. Development Support Tools - Discussion

In parallel to the development of IPL-Asia, we have also developed two software tools: a resource selection support tool and a metadata editor based on the workflow shown below. The workflow is an elaboration of steps 1 and 2 of the process model described in Section 3.3.

Stage 1. A person, called a contributor, who finds a high quality resource deposits a minimum set of information into a database.

Stage 2. A cataloger who understands the language of the resource creates metadata for the chosen resource.

Stage 3. The cataloger communicates with the contributor and tries to get sufficient information to start writing metadata in the language of the resource, and in the common language used among the group of catalogers, i.e. Japanese in IPL-Asia.

Stage 4. The group of catalogers collaboratively edits the metadata and adds it to the database for access.

This paper does not describe the tools in detail, but can be seen in [11].

4. Lessons Learned

4.1. Criteria for Internet Resources Selection

The criteria for Internet resource selection are composed of four aspects: "Accessibility", "Usability", "Reliability" and "Currency".

(1) Accessibility

- Readability of the contents. Pages with too many banners or advertisements should be ranked low.
- Visibility of access policy, i.e., free or charged. If a resource is not accessible freely, the charging policy should be explicitly stated.
- "Navigatability" of a resource.

(2) Usability

- Visibility of subjects and audience.
- Stability of contents and accessibility.
- Appropriateness of contents of a resource to its target audience.
- Educational and emotional quality of a resource for its target audience.

(3) Reliability

- Source and/or publisher of the resources. Resources published by authoritative organizations should be ranked high.
- Identifiability of a creator(s) of a resource.

(4) Maintenance Quality

- Update frequency.
- "Up-to-dateness" of a resource.

Contributors, who search and collect appropriate resources, are given a set of written guidelines based on these criteria. (In this prototype development, the participants who worked as contributors also cataloged the resources themselves).

These criteria and associated guidelines are useful for the contributors to ensure consistency of resource selection and hence, service quality. However, in a practical environment, Internet resources are not necessarily evaluated strictly in all of these criteria.

4.2. Metadata Schema and its Documentation

Our framework for metadata creation – which is to create metadata in multiple languages by translation – has been a challenging task. The key issues to perform this task included not only the definition of the metadata schema and its guidelines but also its documentation.

We had to evaluate the metadata schema and its documentation from two viewpoints: language-neutral applicability to resources and usability of the schema in language-dependent descriptions. As shown in Table 1, the meaning of any element is language-neutral although some element values are expressed in a natural language or in a language-dependent vocabulary. In the design of the schema, we made efforts to minimize the number of elements whose description is language dependent. For example, Title and Description are obviously language-dependent. The Subject element is basically a language-neutral element but its value would include culture-specific and location-specific terms.

Documentation of the schema and its associated guidelines was another key in the prototype development. The documents for the schema and the guidelines were prepared in Japanese only, which was the common language among the participants who physically shared the working environment in this project. We did not have problems in relation to understanding the Japanese documents but there were times face-to-face consultations are necessary because of language differences. Translation of the documents into Chinese and Korean is required to make the task easier and globally distributed among voluntary catalogers collaborating via the Internet.

We also found that translating metadata description and writing metadata from scratch are entirely two different tasks. A cataloger who translates metadata description does not necessarily understand the language of the object resource, which means that metadata description in an original language has to provide sufficient information for the translator. Moreover, another issue in translation concern transcription of terms, especially Japanese names. However, this is not specific to this system.

4.3. Support for Catalogers

Since metadata records were created mainly by graduate students at a laboratory, they were able to have face-to-face meetings when problems arose.

Training and orientation for catalogers are an important issue, which is easily taken for granted in a research environment. The basic business model of IPL-Asia is collaboration by volunteer catalogers. Since the prototype development was performed in an environment where in-person communication is easy, scaling up the task in the Internet environment while maintaining metadata quality would rely heavily on training and orientation resources for catalogers. The Open Directory Project [16], a large-scale collaboration by volunteer editors through the Internet, is a good example in this respect

4.4. Metadata Creation and Evaluation

We created metadata records for two types of resources: Internet resources for children and the digital collection provided by Japanese public libraries which are

mainly created by digitizing precious historical materials. The former includes the resources expressed in CJK, but the latter includes only Japanese materials.

We have not evaluated the costs involved to create multilingual metadata. Experiences in this experiment roughly showed that it took about three to five hours to create a metadata record in CJK languages for Internet resources. It took longer to create metadata for the public library resources because of the nature of the resources. Catalogers had to cope with historical and local terms which are hard to understand and whose proper pronunciation is hard to find.

Collaborative metadata creation turned out to be a more time-consuming task than was expected. Although human catalogers are expensive, they are considered indispensable for high quality description of metadata. Cooperation among specialists (librarians, translators) would be an ideal environment. However, cooperation among volunteers would also suffice as this experiment had demonstrated. Software tools supporting collaborative work are essential to decrease the cost in distributed environment.

4.5. Requirements to Software tools supporting collaborative metadata creation

All metadata was encoded in XML. XML was an appropriate choice to organize user interfaces in multiple languages and to perform access functions to metadata records. Since there was no off-the-shelf tool applicable to edit and organize XML-based multilingual metadata records when we started this study, we had to start metadata creation using a simple editor and basic XML tools. We made the schema and guideline documents available on the Web for catalogers. This simple environment worked but caused some inefficiency for collaborative metadata creation. We were able to develop a metadata creation process model, which was reflected on the design of the support tools mentioned in Section 3.5.

We did not use machine translation software for this project. However, evaluation of the feasibility of using machine translation software to support catalogers is left for future work.

4.6. Adaptable User Interface

IPL-Asia provides user interfaces in CJK languages and English. Users can choose any single language. Metadata records can be displayed either in a single language or in multiple languages in accordance with user's preference.

IPL-Asia provides a directory-style interface. The directory is organized according to a classification scheme. We have found that the directory could be made adaptable in accordance with user age aside from user language preference. Since it is too expensive to prepare separate metadata databases in accordance with user age,

we need to build an adaptable directory-style user interface to the database in accordance with user age.

Ontology technologies developed in the context of Semantic Web activities should be evaluated for further development. Those technologies will be useful to create adaptable user interfaces and for maintenance of metadata vocabularies.

5. Future Work

Collaboration among participants over the Internet is important in the development of subject gateways. Software tools to support multilingual collaborative development of metadata are crucial for IPL-Asia. We need to evaluate the functionality of the software tools experimentally developed in practical working environments.

In addition, metadata vocabularies for a multilingual subject gateway and software tools to use the vocabularies are left for future work. For example, subject terms, which are assigned based on language-neutral vocabulary and can be expressed in all of the cataloging languages, are very important because a user can get information from a metadata record even if not translated into one's language

6. Conclusions

Creating a metadata record for a single information resource in four languages was a big challenge. Metadata in multiple languages is interesting and useful in the global information environment.

In this experimental study, catalogers are neither professional catalogers nor translators. Although cataloging task was more time-consuming than expected, we think that a volunteer-based collaboration model is effective to decrease costs and to cover a wide range of resources. In this study, graduate students collected resources for children and assigned metadata. We would like to create an environment where children can also actively participate most likely in the area of selection of resources for their own use.

Acknowledgements

We would like to acknowledge Qiang Zhao and Aiko Kato for their contributions to this study. We would also like to thank all ULIS students who participated in this study.

References

- [1] Schäuble, P., Smeaton, A.F. (eds), An International Research Agenda for Digital Libraries – Summary Report of the Series of Joint NSF-EU Working Groups on Future Directions for Digital Libraries Research, October 1998. <http://www.iei.pi.cnr.it/DELOS/NSF/Brussrep.htm>
- [2] Internet Public Library (IPL). <http://www.ipl.org>

- [3] AHDS (Arts and Humanities Research Board). <http://ahds.ac.uk/>
- [4] BIOME. <http://biome.ac.uk/>
- [5] SOSIG (Social Science Information Gateway). <http://www.sosig.ac.uk/>
- [6] Biz/ed. <http://www.bized.ac.uk/>
- [7] GEM. (Gateway to Educational Materials) <http://thegateway.org/>
- [8] RDN (Resource Discovery Network). <http://www.rdn.ac.uk/>
- [9] Renardus. <http://www.renardus.org/>
- [10] WWW Virtual Library. <http://www.vlib.org.uk/>
- [11] Shigeo SUGIMOTO, et al., An Experimental Study on a Collaborative Development of a Multilingual Subject Gateway for Public Library Users, Proceedings of Digital Library – IT Opportunities and Challenges in the New Millennium, pp.118-130, July 2002, Beijing, China
- [12] Japanese Old Tales. <http://www.dl.ulis.ac.jp/oldtales>
- [13] Akira Maeda, et al., Multilingual Documents on Your Local Web Browser, Communications of the ACM, vol.41, no.4, pp.64-65, April 1998.
- [14] Dublin Core Metadata initiative. <http://dublincore.org/>
- [15] IEEE Learning Technology Standards Committee. <http://Itsc.ieee.org/>
- [16] Open Directory Project. <http://dmoz.org/>