DCPAPERS

Toward Core Subject Vocabularies for Community-oriented Subject Gateways

Wonsook Lee
Graduate School of Library,
Information and Media Studies
University of Tsukuba
Tsukuba, Ibaraki, Japan
wonsook@slis.tsukuba.ac.jp

Shigeo Sugimoto
Graduate School of Library,
Information and Media Studies
University of Tsukuba
Tsukuba, Ibaraki, Japan
sugimoto@slis.tsukuba.ac.jp

Abstract

Subject classification schemes and vocabularies for subject indexing, i.e. subject vocabularies, have a crucial role in the development of subject gateways. From our experiences on a few subject gateway projects which are designed for regional and domain-specific communities, we learned that the subject gateways require a subject vocabulary which is reasonably small and tailored in accordance with the resources in the domain and the audience.

The goal of this paper is to discuss underlying issues for the small "core" subject vocabularies defined as a subject description/classification scheme for community oriented resources. First, this paper describes some basic aspects for metadata schema and requirements for the region/domain-specific subject gateways. Second, we show three subject gateway projects which are resources designed primarily for libraries and public library users. And then, we show a few examples of small vocabularies designed for regional resources. This paper compares some vocabularies designed for regional resources and discusses issues for the core subject vocabularies.

Costs to develop and maintain the vocabularies can not be neglected even if they are small. We consider the Semantic Web technologies will help develop software tools to support development and maintenance of the subject vocabularies and to enhance their interoperability and reusability.

Keywords

Subject Gateway, Community-oriented Subject Vocabulary, Core Vocabulary, Interoperability

1. Introduction

Vocabularies for classification of resources and those for expressing subjects are a crucial component of a subject gateway which organizes information resources in accordance with the domain and the audience. Traditionally, large-scale controlled subject vocabularies have been widely used in the library community, e.g. Dewey Decimal Classification (DDC), Nippon Decimal Classification (NDC), Library of Congress Subject Headings (LCSH) and so forth. These well-established vocabularies are often used as a subject classification scheme by subject gateways. However, on the other hand, a subject gateway which is designed for a community built based on a specific domain or a region needs a subject vocabulary tailored to the community, i.e. the domain and the audience of the community.

The authors have been involved in a few subject gateway projects – a digital library service at University of Library and Information Science (ULIS-DL)[1][2], IPL-Asia[3] and Digital Okayama Dai-Hyakka[2]. These projects are primarily planned as a subject gateway. Dublin Core is a common metadata schema for these projects. We learned a few common issues critical for the development of these projects, e.g. schemes and vocabularies to classify resources, technology and methodology to efficiently find and collect valuable resources, and long term maintenance of the metadata records and schemas.

ULIS-DL which started in early 1999 has collected

https://doi.org/10.23106/dcmi.952107999

resources published by libraries and those of Library-and-Information-Science (LIS) domain. The metadata schema of ULIS-DL is defined based on Simple DC with a few extensions. When we started ULIS-DL we did not employ any controlled vocabulary to describe subjects or genre of a resource. The main reason was that we could not find an appropriate subject vocabulary for ULIS-DL which is aimed to collect resources from specific type of resources, i.e. homepages of libraries and LIS-related institutions/organizations. In 2003, we examined the terms given as subject terms, i.e. texts given in Subject element. We learned a significantly large part of the metadata records can be covered by a relatively small set of subject terms from this study.

IPL-Asia is an effort to provide metadata of resources useful public library users which is expressed in multiple languages - Chinese, Japanese and Korean (CJK). The metadata schema of IPL-Asia is defined based on Dublin Core and Learning Object Metadata (LOM). In IPL-Asia, We adopted UDC as the subject vocabulary of IPL-Asia. 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. For example, we need to choose terms in accordance with age ranges of expected audience and each term should be represented in an appropriate style in accordance with the age ranges.

Digital Okayama Dai-Hyakka (DODH) is a regional portal by the Okayama Prefecture Library (OPL) in collaboration with other public sectors in the prefecture of Okayama. DODH provides information about regional resources published on the Web. The metadata schema of DODH is defined based on Simple Dublin Core with a few extensions. DODH has three subject vocabularies, which are NDC, a subject vocabulary defined for resources published by the prefecture government, and a subject vocabulary for children (KV). KV was designed based on the experiences in IPL-Asia – it has approximately 300 terms and each term has labels defined in accordance with age ranges. A crosswalk table, which is a mapping table among the three vocabularies, was defined on top of the vocabularies by OPL.

From these case studies, we have learned;

A reasonably small subject vocabulary which is built in accordance with the application domains and the audience is preferable rather than well-established large subject vocabularies.

Software tools to enhance interoperability between the subject vocabularies and to improve maintainability and re-usability (or customizability) of the subject vocabularies are seriously required.

Based on these lessons learned, in this paper, we examined the characteristics of some subject vocabularies developed for regional resources.

In the rest of this paper, we describe some discussions on metadata schemas (section 2), three case studies (section 3), and discussions on based on the case studies (section 4).

2. Metadata schemas and vocabularies for subject gateways

2.1 Some Definitions

Before going into detailed discussion, this section gives a set of brief descriptions of technical terms used in this paper.

Metadata schema: A metadata schema is a description scheme of metadata. Generally, metadata schema consists of definitions of metadata terms to express properties and those to express classes of values, definitions of syntactic constructs and constraints, and implementation syntax.

Metadata term: An atomic entity included in a metadata schema. A term represents a property of a resource or a class of values. A value could be a resource or a literal value. A term is given a machine readable unique identifier and can be assigned one or more labels for human understandability.

Metadata Element Set: A set of metadata terms included in a metadata schema. For example, DCMES is an application neutral metadata element set. An application specific element set can be defined by extracting metadata terms from other element sets.

Application Profile: An application profile defines a set of structural and syntactic constraints applied on a set of metadata elements[4]. Metadata elements and an application profile(s) are a part of a metadata schema.

Metadata vocabulary: A metadata vocabulary is a set of metadata terms. The meanings of metadata vocabulary and metadata element set are the same in this paper.

Property vocabulary: A set of property terms. A property term is a metadata term which represents a resource property, i.e., property term. A property term, or simply property is a term which expresses a property (or attribute) of a resources. A property value of a resource is a value associated by a property.

Class vocabulary: A set of class terms. A class term is a metadata term which represents a class of resources or a class of literal values. A class term, or simply a class, is a term which expresses a set of property values or a scheme to express property values. For example, “DCMI Type Vocabulary” is a class term.

Value vocabulary: A set of value terms. A value term is a metadata term which represents a value instance. For example, “collection” of “DCMI Type Vocabulary” is a value term.

https://doi.org/10.23106/dcmi.952107999

Subject vocabulary: A set of subject terms. A subject term is a value term which expresses a subject, i.e., subject category or classification. For example, DDC, NDC, and LCSH are subject vocabularies. Thus, a subject vocabulary is a class vocabulary.

2.2 Background

Subject classification schemes and subject headings, which are called subject vocabularies, are a key component for subject gateways to organize information resources and build resource indexes for search and navigation functions. Well-established classification schemes such as DDC and NDC have been used in subject gateways for scholarly information resources, e.g. Renardus [5], JuNii [6].

Generally, those subject vocabularies, which have been widely used in the library community, broadly cover scholarly domains and are primarily designed for traditional library resources, e.g., books, periodicals, scholarly articles, and so forth. On the other hand, the subject gateways for narrower domains and/or for specific audience need to have subject classification schemes in accordance with the domains and audience, e.g. subject gateways for regional resources and those for children's resources. For example, the Internet Public Library [7] has the categorization scheme for children's resources which is different from the scheme for general resources.

An application specific subject vocabulary has both advantages and disadvantage.

Advantages:
- It provides a set of terms organized in accordance with the application and the audience.
- It can cover the domain with the appropriate depth.

Disadvantages:
- It may lose interoperability with other applications.
- Maintenance of the vocabulary is expensive.

Maintenance is another key issue for subject vocabularies. However, in general, it is not easy for small communities to pay large cost to maintain a subject vocabulary. This means that a small community has two alternatives, to use well-established comprehensive vocabularies or to build a reasonably small subject vocabulary in accordance with their requirements.

Based on these thoughts, we got a hypothesis that a domain-specific subject vocabulary which is reasonably small is preferable than a large comprehensive vocabulary for communities to build a subject gateway for the community resources. In this paper, the reasonably small vocabulary which is defined based on requirements of a community is called a core subject vocabulary.

2.3 Requirements for Community-oriented Core Subject Vocabulary

In this section, we examine issues required for community-oriented core vocabularies in two aspects; (1) vocabulary size and structure, and (2) interoperability maintenance and reuse.

(1) Size and structure of vocabulary

"Reasonably Small" does not mean a definite upper limit of number of terms in a subject vocabulary or a definite depth of hierarchy of classification. From the experiment of ULIS-DL shown in section 3.1 a set of some 1000 terms covers 90% of 26,000 metadata records. A vocabulary created for resources published by a prefecture government shown in section 3.3 includes about 300 terms. Major part of the decimal classification systems, e.g. DDC and NDC, includes 10x10x10 terms in principle.

From a viewpoint of user interface of resource indexes, it is useless to show too many index terms on a single screen. BUBL[8] has the 5:15 rule which means the number of links from a term to the lower layer terms should be in principle between 5 and 15 with some exceptions. An index page built based on a decimal classification scheme basically contains 10 branches, i.e. 10 subject terms. This feature could be easily applied to form a user interface. However, resources of community-oriented subject gateways will not be distributed equally to all subject areas. In the case of subject gateways for children the index terms in a user interface have to be expressed in a user-friendly form for children. A term could be expressed not only in a textual form but also as an image.

(2) Interoperability, Maintenance and Reuse

Interoperability between subject vocabularies is a crucial issue for cross-community resource discovery. However, it is hard to get a general solution for the issue. A non-technical but feasible way to solve this issue is to manually build a mapping function between two (or more) vocabularies, e.g., crosswalks and switching languages. However, it is very expensive to build a mapping function between large vocabularies. From this viewpoint, a reasonably small vocabulary would be useful as a vocabulary to describe subjects of a resource. A small vocabulary could be used as a switching language to crosswalk or merge more than one subject vocabulary

Long-term maintenance of a vocabulary is a crucial issue. Meaning and usage of a subject term change over time. It is crucial to use an appropriate technology to maintain the vocabulary. Maintenance

https://doi.org/10.23106/dcmi.952107999

of a vocabulary could include versioning of definition of each term, merge and split of terms.

Neighboring communities would be able to share a subject vocabulary but each community would prefer to use its own vocabulary. In that case, because creating a subject vocabulary is an expensive task, it is advantageous to reuse an existing vocabulary and customize it for a new application.

It is obvious that we need software tools to support the three issues - interoperability, maintenance and reuse. XML-based scheme to encode vocabularies is a crucial aspect to realize the software tools. RDF Vocabulary Description Language (RDF Schema) and Web Ontology Language (OWL) are natural candidates as the encoding scheme for subject vocabularies. Technology to provide the vocabularies via the Internet is also crucial. Metadata schema registries have a large potential for this issue. They can provide functions to register a vocabulary and its terms, to search and browse terms and vocabularies stored in a registry or registries, and so forth.

3. Preceding Case Studies on Community-oriented Subject Gateways

3.1 ULIS-DL [1][2]

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 by libraries and LIS institutions mainly in Japan, and created metadata 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. This set is called a candidate term set, CTS-n (n>1). Table 1 shows CTS-2, 3, 4 and 5. The "uncoverage" ratio in Table 1 is defined as the ratio between a number of metadata records that do not have any of the core subject terms and the total number of metadata records.

Table 1. Coverage of Candidate Term Sets
Total number of metadata records = 26358
Total number of Subject element values in the primary term set = 28797
Total number of Subject element values excluding Publisher/Creator/Contributor values = 26107

	number of Subject terms	number of excluded records	Uncoverage ratio
CTS-2	3979	1519	6%
CTS-3	2045	2083	8%
CTS-4	1366	2590	10%
CTS-5	1025	2801	11%

We chose CTS-5 as the core subject vocabulary in this study because it covers approximately 90% of the total records, and also because we considered its size as the most reasonable to manually organize the subject vocabulary terms, i.e. classifying the terms and defining relationships between the terms.

We classified the CTS-5 terms 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 detailed 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. Every subject term of CTS-5 was associated to a leaf node as an occurrence term in the metadata. We encoded the terms and their tree structure in OWL.

3. 2 Internet Public Library Asia

Internet Public Library Asia started in the year 2000 at the University of Library and Information Science (ULIS) in Tsukuba, which was initially planned as a collaborative activity with the University of Michigan in part. This project is reported in detail at the DC-2003 conference in Seattle [3].

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 which are chosen from those of the Dublin Core Metadata Element Set (DCMES) and

https://doi.org/10.23106/dcmi.952107999

Figure 1. Example metadata record described in CJK

IEEE LOM. Each metadata record was collaboratively created by a group of catalogers.

The user interface of IPL-Asia was prepared in multiple languages as shown in Figure 1. 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. Figure 2 shows a part of an example metadata record encoded in XML.

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.

```
<title>
<main-title>
<langstring contents="original" xml:lang="ko">
____ __</langstring>
<langstring contents="translation" xml:lang="ja">
_____</langstring>
  <langstringcontents="translation" xml:lang="zh">
   _____</langstring>
<langstring contents="translation" xml:lang="en">
The characteristic of adolescence </langstring>
</main-title>
</title>
```

Figure 2. XML Description in CJK

3.3 Digital Okayama Dai-Hyakka (DODH) – a Regional Portal by a Public Library

Digital Okayama Dai-Hyakka (DODH) is a regional portal by the Okayama Prefecture Library in collaboration with other public sectors in the prefecture [2][9]. DODH provides a Z39.50-based OPAC across public libraries in the prefecture of Okayama, a reference database, a 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

https://doi.org/10.23106/dcmi.952107999

published by the prefecture government and NDC. The first scheme, which is called Okayama Kids Vocabulary (KV), is designed primarily for 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 prefecture government. Both of these regional subject vocabularies are designed to be sufficiently simple because 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. Mapping tables between these subject vocabularies were created by OPL.

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 for the youngest age group has to be readable and understandable for children of that age, so that we can use only syllabic characters (i.e., Hiragana and Katakana) and a limited set of Kanji (Chinese characters) and we have to re-phrase subject terms into plain words or phrases to make it easier for children to understand. For higher graders of elementary schools, we also used easier words and phrases for presentation labels, and we add pronunciation scripts to the subject terms expressed. Presentation labels designed for junior high school students are almost the same as those for general public but some of them are re-phrased and there is additional pronunciation information (Figure 3).

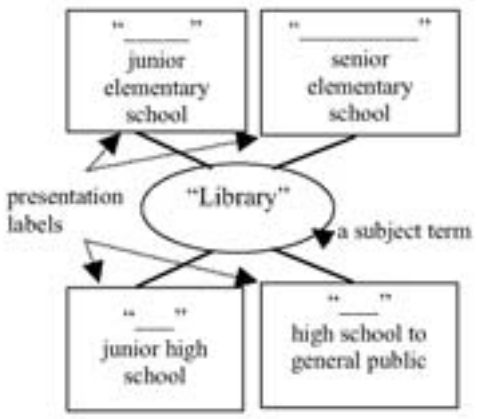

Figure 3. A subject term with multiple presentation labels

4. Discussions on Subject Vocabularies for the Community-oriented Subject Gateways

4.1 Re-examination of Requirements for Small Subject Vocabularies

From the experiences in ULIS-DL and IPL-Asia, we learned that a reasonably small set of terms, which is chosen in accordance with regional/domain-specific resources and with audience and is called a core subject vocabulary, is preferable as a subject vocabulary for the resources than conventional comprehensive and authoritative classification schemes such as DDC and NDC. These conventional subject vocabularies are advantageous for interoperability and stability because they are broadly used in the library community. Their disadvantage is that they are not easily customizable and not fully

used by regional/domain-specific applications. On the other hand, application specific subject vocabularies are expensive to create and maintain but they are tailored for the application and the audience.

A core subject vocabulary for regional resources can be created by designing from scratch or by customizing an existing vocabulary. As described in section 2, the issue of subject vocabulary has the dilemma of preciseness and comprehensiveness of terms vs. simplicity for amateur and child users, interoperability vs. domain-specificity. In the libraries, comprehensive vocabularies have been widely adopted since libraries cover broad range of subjects. However, subject gateways for regional resources have requirements as follows:

It is not easy for amateur users to fully use a large subject vocabulary.

Children sometimes have difficulties to understand terms and characters.

Community specific terms are likely to be excluded from general comprehensive subject vocabularies.

In the ULIS-DL study, we analyzed existing metadata and got a reasonably small set of terms to cover significantly large part of the whole set of metadata. In the study on the DODH vocabulary, we examined example resources and built a set of subject terms. In both studies, we experimentally encoded the subject terms in RDF/OWL. From these studies we found the following issues are crucial to enhance the usability of vocabularies, software tools to support building and maintaining a subject vocabulary, requirements analysis to enhance reusability/customizability of an existing small vocabulary to another similar application, and requirements analysis to cross-use small vocabularies.

For the first issue, we have experimentally built a software tool to edit and maintain vocabularies based on RDF/OWL [10]. In addition to the evaluation of the

https://doi.org/10.23106/dcmi.952107999

Table 2 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 term. "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 293 terms	#Terms	17	8	8	196	58	54	28	62	6	6	443
	Ratio (%)	3.8	1.81	1.81	44.2	13	12	6.3	14	1.35	1.35	100
	#NDC terms	7	7	3	44	27	27	20	26	4	4	169
PV 287 terms	#terms	15	2	12	171	30	56	44	17	1	1	349
	Ratio (%)	4.3	0.6	3.4	49	8.6	16	13	4.9	0.3	0.3	100
	#NDC terms	4	2	5	34	11	18	25	15	1	1	116

tool, our future works on this tool are, connection of the tool to the metadata schema registry developed at Tsukuba [11], and extension of the functionality for reuse and customization of an existing vocabulary for a new application, which is related to the third issue mentioned above.

The next section describes a case study on the second issue comparing three subject vocabularies used in ORIN. In section 4.3, we show a few subject vocabularies defined to classify prefectural government resources and an experimental development of a core subject vocabulary.

4.2 Comparison of Subject Vocabularies in a Regional Subject Gateway – A Case Study

The regional subject gateway at Okayama prefecture (DODH) uses three subject vocabularies for classification of the resources, which are NDC, Okayama Kid's Vocabulary (KV) and Okayama Prefecture Vocabulary (PV). Mapping tables between the vocabularies were created, i.e., between NDC and KV, NDC and PV, and KV and PV. NDC is a decimal classification system widely used in the Japanese library community. Table 2 shows the distribution of terms of KV and PV in the NDC term space using the mapping tables. Table 3 and 4 show mappings to NDC terms from the terms of major categories of KV and PV, respectively.

In the mappings from KV to NDC and PV to NDC, a single term is mapped to one or more NDC terms and more than one term of KV can be mapped to a single NDC term. As shown in Table 3, 169 and 116 NDC terms were used in KV-to-NDC and PV-to-NDC mappings, respectively. The rows #Terms and Ratio mean the numbers of mappings and their rations from KV/PV terms to terms in each NDC category. The Social Sciences category (300) is the most densely mapped category both in KV-NDC and PV-NDC mappings. The maximum numbers of mappings to a single NDC term from KV and PV are 25 (NDC: 374 (school management)) and 27 (NDC: 317 (administration)), respectively. In KV-NDC and PV-NDC mappings, 56% and 54% of the NDC terms have a single input, respectively. Figure 4 shows an outline of the relationships among the three vocabularies.

Table 3 Terms of Major Categories of KV

Terms of Major Categories	NDC terms
IT (Information Technology)	007, 547
_____ (Culture, Arts)	361, 700
_____ (Entertainment, Sports)	780, 790
__ (School)	374, 375, 376, 377, 378
___ (Life)	590
__ (Society)	300
_____ (Region, Human, History)	200
_____ (Environment, Science)	400, 519

https://doi.org/10.23106/dcmi.952107999

Table 4 Terms of Major Categories of PV

Terms of Major Categories	NDC terms
__ (Education)	370
____ (Human, Human Rights)	289
_____ (Welfare, Social Contribution)	369
_____ (Health, Hygiene)	498
_____ (Life, Security)	317, 590
__ (Environment)	519
_____ (Industry, Economy)	330, 600
__ (Labor)	366
_____ (Science, Culture)	000
_____ (Regional Information, History)	210, 291, 361
_____ (Other Prefectures)	210
_____ (Regional Assembly, Administration, Court)	314, 317, 327
____ (Statistics)	350
____ (Infrastructure Development)	601
__ (Tourism)	689
____ (Public Facilities)	317, 318

4.3 "Core" Subject Vocabulary for Regional Governmental Resources – a Case Study

We collected classification vocabularies from four prefecture governments in Japan in order to compare regional vocabularies and examine a model to cross-use the vocabularies. Each of the vocabularies is created by a prefecture government to classify resources published and made available on the Web by the government. As shown in Table 5, all vocabularies are a small set of classification terms. We could not find a common vocabulary from which the four vocabularies are created. The four vocabularies share several concepts which are expressed in different terms. The structures of the four vocabularies are significantly different even though all of them are designed to classify documents published by public sectors; for example, total numbers of their classification terms are different, numbers of their top level terms are different, and a same term appears in different classification levels, and so on. A common subject vocabulary is required to classify resources across these prefectures. We experimentally merged and re-classified the terms in the four vocabularies and built a "core" subject vocabulary which has 18 top level terms. This vocabulary was created by extracting common concepts that commonly appear in the top

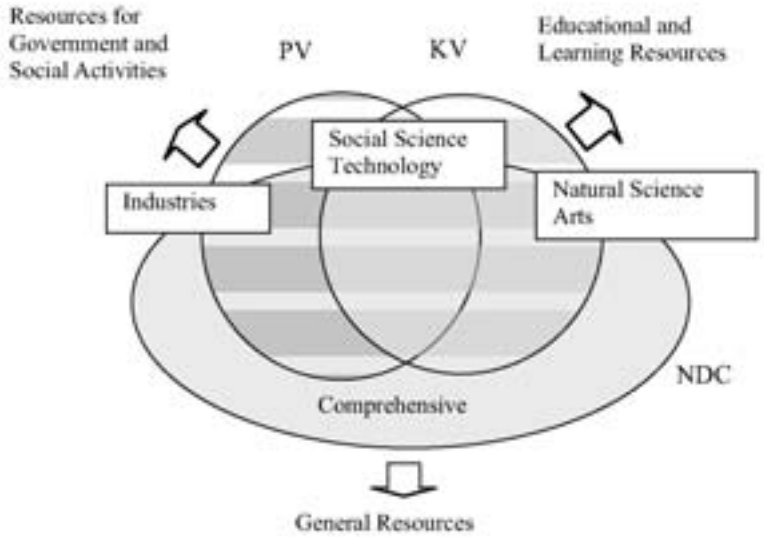

Figure 4 Outline of Coverage of Vocabularies

https://doi.org/10.23106/dcmi.952107999

Table 5 Statistics of Classification Vocabularies of the Four Prefectures

	Ibaraki	Kagawa	Kanagawa	Okayama
Number of Top Level Terms	23	11	14	16
Max Depth	2	2	2	3
Total Number of Terms	115+	82	88	280+

level of the four vocabularies. Each term of the core vocabulary has linkage to corresponding terms of the four vocabularies. Our future work is the evaluation of the core vocabulary, which would include usability as a subject vocabulary across regions and a switching language [12].

5. Conclusion

Well-organized subject vocabularies are a crucial component for subject gateways. They are important not only for the developers to organize information resources but also for the users to find useful resources. The library community has rich and long experiences in developing and using the subject vocabularies to classify and index library resources. The library community has been paying handsomely to maintain subject vocabularies. In the library environment, the subject vocabulary is an "already-established" issue.

On the other hand, in the Web environment, we need to organize various types of information resources, which are different from traditional library resources, and provide them to various types of users via the Internet and off-the-shelf browsers. From our experiences in our subject gateway projects, we learned that a good navigational tool to browse metadata is required and that subject vocabularies are an indispensable component to realize the tool. At the same time, we learned that conventional vocabularies were not suitable and that subject vocabularies which are designed for application domains but need not be large are required for our projects.

Region/domain-specific vocabularies are advantageous to organize resources in accordance with the region/domain-specific requirements but, on the other hand, they would have risks to decrease interoperability between subject gateways, and to deteriorate maintainability and customizability (or reusability) of the vocabularies. We are hoping that ontology technologies for the vocabulary and software tools based on the Semantic Web such as RDF/OWL are crucial to cope with these issues.

Acknowledgements

DODH is a project at Okayama Prefecture Library. The mappings discussed in this paper were created by this library. The authors express our sincere thanks to Mr. Mitsuyoshi Moriyama of Okayama Prefecture Library and his colleagues for their contribution. We express our thanks to all of our colleagues who have contributed to the related projects.

References

1. M. Murotani. Development of a Subject Vocabulary and Directory-style User Interface for ULIS-DL, Master Thesis, Tsukuba, Japan, University of Tsukuba, 2004 (in Japanese)
2. S. Sugimoto, et al. "Developing Community-Oriented Metadata Vocabularies: Some Case Studies" Proceedings of DLKC'04, 2004.3, pp.128-135.
3. WS. Lee, et al. "A Subject Gateway in Multiple Languages: a Prototype Development and Lessons Learned" Proceedings of DC-2003, Seattle, 2003, pp. 59-66.
4. Dublin Core Application Profile Guidelines, CEN Aorkshop Agreement (CWA) 14855, 2003
5. Renardus. http://www.renardus.org/
6. JuNii. http://ju.nii.ac.jp/
7. Internet Public Library, http://www.ipl.org/
8. BUBL. http://bubl.ac.uk/
9. Digital Okayama Dai-Hyakka (DODH). http://www.libnet.pref.okayama.jp/mmhp/index.html (in Japanese)
10. M. Furuta. A Model and a Software Tool for Maintaining Community-oriented Metadata Vocabularies Master Thesis, Tsukuba, Japan, University of Tsukuba, 2005
11. M. Nagamori and S Sugimoto. "A Metadata Schema Framework for Functional Extention of the Metadata Schema Registry" Proceedings of DC-2004, Shanghai, 2004, pp 3-11.
12. Jens-Erik Mai, "The Future of General Classification", Cataloging & Classificaion Quarterly, Vol.37, No1/2, 2003, pp.3-12

https://doi.org/10.23106/dcmi.952107999