Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Providing universal access to Japanese humanities digital libraries: an approach to federated searching system using automatic metadata mapping*

Biligsaikhan BATJARGAL†[1], Fuminori KIMURA[2], Akira MAEDA[2]

([1]*Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan*)

([2]*College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan*)

†E-mail: biligsaikhan@gmail.com

Received Sept. 1, 2010; Revision accepted Oct. 12, 2010; Crosschecked Sept. 1, 2010

Abstract: We present our approach for constructing a federated searching system for Japanese humanities digital libraries using automatic metadata mapping. We discuss some achievements in the ongoing research to construct a federated searching system. The goals of this system are (1) to perform metadata mapping automatically for Japanese heterogeneous humanities digital libraries and (2) to let users access multiple humanities digital libraries by using only one query input. The proposed automatic metadata mapping method produced the average mapping precision of 94.9%. We also address the metadata-related challenges facing Japanese humanities databases.

Key words: Federated searching system, Metadata mapping, Humanities digital libraries, Universal access
doi:10.1631/jzus.C1001001 **Document code:** A **CLC number:** TP391

1 Introduction

In recent years, the role and importance of providing universal access to information has continuously increased. Increasing public demand encourages digital library systems to be multilingual and to support various collections in different languages. Many humanities collections and resources in libraries, museums, and research institutes are digitized and made available for public viewing in a wide variety of languages. In library and information science, metadata is a vital solution for describing and managing the massive quantities of an explosively growing, complex world of digital information (Zeng and Jian, 2008). Various types of resources and humanities digital libraries coexist with heterogeneous metadata schemas nowadays, and many different metadata schemas are standardized by international standards organizations.

How to deal with the diverse forms of metadata and their interoperability is becoming a complex issue for research. There have been efforts to make heterogeneous standards interoperable and use multiple metadata standards. According to Chan and Zeng (2006), several different approaches (element mapping, crosswalk, application profile, metadata registry, etc.) were developed. Reliable metadata interoperability has not been achieved yet because of the heterogeneity of metadata standards and because of the structural differences between standards.

On the other hand, the use of metadata schemas and standards for Japanese humanities digital libraries is a bit tricky. Digital contents in Japanese could be one of the biggest non-English and non-western content in the world. Many metadata schemas of Japanese humanities digital libraries have been accepted in terms of their semantics and content but

* Project supported by the Grant-in-Aid for the Global COE Program from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, and MEXT Grant-in-Aid for Strategic Formation of Research Infrastructure for Private University, Japan

were developed before the international metadata standards or were developed without considering the international metadata standards or specific encoding methods. Most of the metadata schemas of Japanese humanities digital libraries were not derived from existing international metadata standards, and there is no explicit metadata framework, crosswalk, or metadata registry. It is necessary to understand the semantics of Japanese humanities digital libraries—such as elements, syntax, and structure—in order to perform automatic metadata mapping and achieve metadata interoperability. This paper therefore addresses the metadata-related challenges for the Japanese humanities digital libraries. Moreover, we present our approach for constructing a federated searching system for Japanese humanities digital libraries, using automatic metadata mapping.

2 Metadata schemas for Japanese humanities digital libraries and their challenges

Humanities digital libraries and their metadata schemas are very heterogeneous because the humanities cover a variety of disciplines, such as literature, law, history, philosophy, religion, visual and performing arts (including music), anthropology, cultural studies, and linguistics (including ancient and modern languages). Achieving metadata interoperability of humanities digital libraries is becoming more crucial in the current information environment, especially in the case of metadata schemas which were not derived from well-known international metadata standards.

One of the differences between western and Japanese databases that are relevant to people interested in constructing a federated searching system is the greater heterogeneity of the metadata schemas of Japanese humanities digital libraries.

Many Japanese humanities databases developed metadata schemas based on their domain-specific semantics and content rather than adopted international metadata standards. Moreover, names or labels for metadata attributes/elements are written in Japanese, or labels in Japanese are used as the metadata elements. The co-existence of non-standard and heterogeneous metadata schemas makes automatic metadata mapping for Japanese humanities databases a rather challenging task.

Another relevant difference is the Japanese writing system(s). Japanese is written in a mixture of three writing systems—one using ideographic symbols, or 'kanji', and the other two using the syllabary scripts 'hiragana' and 'katakana'—and it is written without explicit word boundaries. The absence of word delimiters makes word segmentation (i.e., tokenization) a critical problem in natural language processing for Japanese. Without knowing the boundaries of words in a sentence, any computer system will fail to perform tasks such as automatic metadata mapping. A single kanji can have many pronunciations and be used differently in words comprising two or more kanji. The situation will be much more difficult when collections contain ancient documents because a modern kanji is not always the same as its archaic equivalent. An archaic word written with a single kanji might be equivalent to a modern word written with more than a single modern kanji, or vice versa. Using a modern language query to find information in Japanese documents that are written in modern and archaic Japanese words is a rather challenging task.

Recently, the National Institute for the Humanities of Japan developed the Resource Sharing System for Humanities (http://www.nihu.jp/sougou/kyoyuka/tougou/index.html) that performs federated searching across the humanities digital libraries of seven organizations, including the National Museum of Japanese History, National Institute of Japanese Literature, National Institute for Japanese Language and Linguistics, International Research Center for Japanese Studies, Research Institute for Humanity and Nature, and National Museum of Ethnology. Hundreds of metadata schemas of Japanese humanities digital libraries were mapped manually to the Dublin Core Metadata Element Set (DCMES) (Dublin Core Metadata Initiative, 2008) and indeed, it was time-consuming and costly. Thus, we developed the automatic metadata mapping method (Kimura *et al*., 2009).

3 Federated searching system for Japanese humanities digital libraries

Constructing a federated searching system for humanities digital libraries is the goal of our ongoing research. The conceptual architecture of our proposed

federated searching system is shown in Fig 1. As illustrated there, if a user wants to find a humanities resource with the query word in the title, our system retrieves resources having the query word in the title or any metadata field that is similar to a title or could be treated as a title, and retrieves these resources from heterogeneous humanities digital libraries even if those libraries do not provide metadata interoperability or crosswalk and do not support the Z39.50 protocol, search/retrieve Web service (SRW)/search/retrieve via URL (SRU), etc. We used the automatic metadata mapping method of Kimura *et al*. (2009). In our system the metadata attribute names of heterogeneous Japanese humanities collections in Japanese, the metadata schemas of which are unknown or do not conform to the international standards, are automatically mapped to our modified variant set (hereafter, modified DCMES) of the Dublin Core Metadata Element Set (DCMES) (Dublin Core Metadata Initiative, 2008). Because CREATOR and CONTRIBUTOR are hard to distinguish in Japanese humanities collections, in the modified DCMES, they are unified into the new element AUTHOR. When Japanese humanities metadata schemas are successfully mapped to the modified DCMES, our proposed system enables cross-domain metadata harvesting and federated searches as well as the exchange of metadata. Our automatic metadata mapping method successfully mapped 334 attribute names of Japanese humanities collections to metadata elements of the modified DCMES with an average mapping precision of 94.9% (Kimura *et al*., 2009).

The next section discusses our achievements on federated search and approach to (re)using existing metadata elements for other purposes without omitting them.

4 Implementation, evaluation, and experiments

Based on the achievements of our automatic metadata mapping method, we have started to construct a federated searching system and conducted an experiment to retrieve information from heterogeneous Japanese humanities digital libraries. In an experiment performing a federated search using automatic metadata mapping, we used Japanese humanities databases—including the image database of Japanese traditional fine art Ukiyo-e (http://www.dh-jac.net/db/arcnishikie/searchp.htm) (67 metadata elements), the donated Japanese books database (http://www.dh-jac.net/db3/Books/default.htm) (11 metadata elements), and the old Japanese books database (http://www.dh-jac.net/db1/books/search.html) (21 metadata elements)—that are freely accessible in Japanese at the Art Research Center of Ritsumeikan University (Kawashima *et al*., 2009; Akama *et al*., 2010).

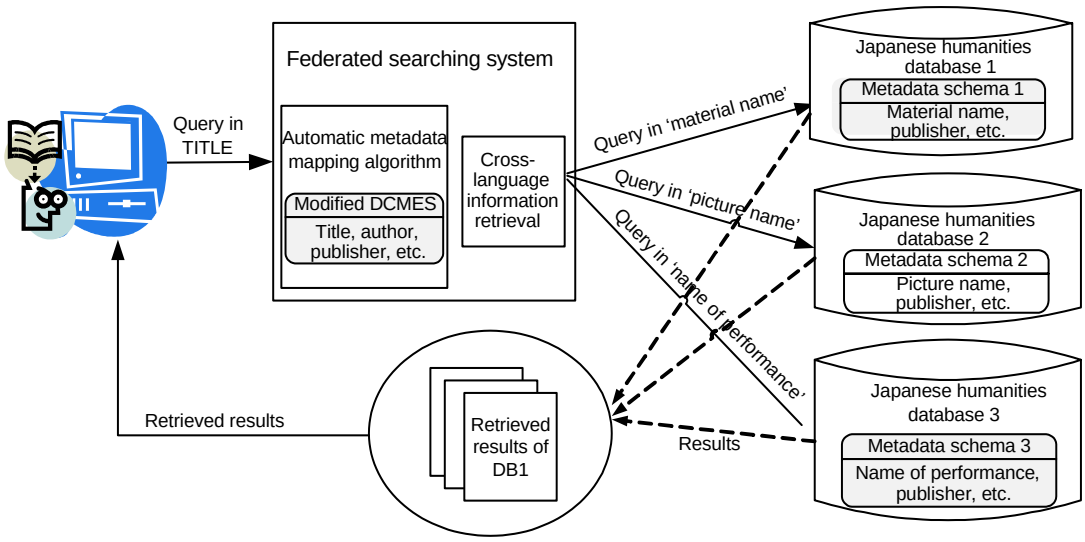

Fig. 1 Conceptual architecture of the proposed federated searching system

4.1 Automatic metadata mapping

Our automatic metadata mapping method (Fig. 2) consists of two preprocessing phases and four mapping phases.

The preprocessing consists of the following steps:

P-1: Collect attribute names from humanities databases for training and mapping.

P-2: Classify attribute names for training into appropriate metadata elements manually.

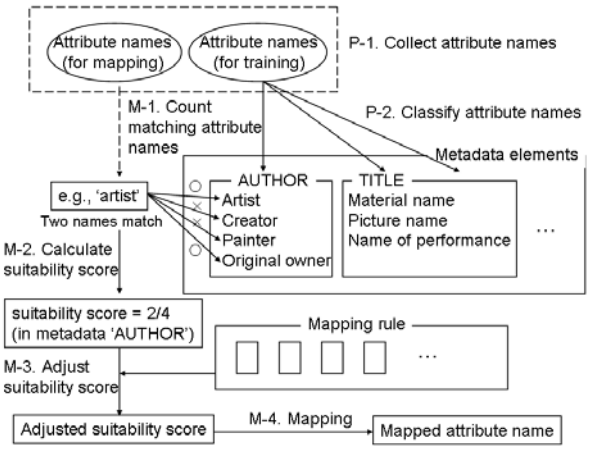

Fig. 2 Flow of automatic metadata mapping

The automatic mapping phase consists of the following steps:

M-1: Count the number of partial string matches between the attribute name for mapping and each metadata element.

M-2: Calculate the suitability score of each metadata element by dividing the number of partial string matches by the number of attribute names in the metadata element.

M-3: Adjust the suitability score for each metadata element, if the target attribute name matches one or more mapping rules, which consist of some kanji characters (or partial words) that are commonly used and known to be relevant to one or more particular metadata elements (e.g., increase the suitability score for 'TEMPORAL' if the attribute name includes 'year').

M-4: Map the target attribute name to the metadata element that has the highest suitability score.

If the attribute name is given the suitability score value 0 for all metadata sets, the attribute name is classified into 'OTHER' metadata.

Inspecting the data listed in Table 1, one sees that 18 metadata elements (attribute names) of the Ukiyo-e image database, donated books database, and old books database were mapped to the TITLE

Table 1 Example of results of the automatic metadata mapping.

Modified DCMES elements mapping	Metadata elements of Japanese humanities digital libraries	Meanings of kanji used in metadata elements of Japanese humanities digital libraries	Number of elements
TITLE	画題等, 画題2, 役名, 外題, 外題よみ, 所作題, 所作題よみ, 細目題, 細目題よみ, 主外題, 主外題よみ, 系統分類題, 演目(統合), 演目よみ（統合）, 画題統合, 資料名, 資料名よみ, 解題	Print title, Picture name, Character names / Official title, Played title / Title of play, Reading of played title, Performed title, Reading of performed title, Detailed title, Reading of detailed title, Main performed title, Reading of the main performed title, Classification title, Name of performance, Reading of the performance, Title of the integrated picture, Material name, Reading of material name, Synopsis	18
PUBLISHER	版元文字, 異版, 版印 1, 版元 1, 版元名 1 , 版印 2, 版元 2, 版元名 2, 版元備考, 地域版, 版元統合	Character publisher, Different edition, Edition stamp1, Publisher 1, Publisher name 1, Edition stamp 2, Publisher 2, Publisher name 2, Publisher remarks, Domestic publisher, Joint publisher	11
DATE	西暦, 和暦, 年月日備考, 月日計算, 西暦版, 和暦版, 月日版, 年月日備考版, 閏, 月, 日	Gregorian calendar, Japanese calendar, Edited date, Date calculation, Gregorian calendar edition, Japanese calendar edition, Edition date, Remarked date, Intercalary, Month, Day	11
AUTHOR	絵師, 編著者等, 原所蔵者, 彫師等, 担当者	Artist, Volume author etc., Original owner, Engravers, etc., Person in charge	5
COVERAGE	地域, 位置, 続方向, 劇場, 場立, 場名	Performed Place, Location, Spatial, Theater, Place, Place name	6

DCMES: Dublin Core Metadata Element Set

element in the modified DCMES. Similarly, 11 elements were mapped to DATE, 11 to PUBLISHER, 5 to AUTHOR, and 6 to COVERAGE. These 18 attribute names were written in various kanji characters that have different meanings, such as 'Print title', 'Picture name', 'Character names', 'Official title', 'Played title', 'Title of play', 'Reading of played title', and 'Performed title'. The metadata attribute names used in Japanese humanities digital libraries consist of several words that have combinations of single or several kanji characters, and the meaning of the words depends on the combinations. Our algorithm performs automatic mapping by calculating the overall suitability scores for each metadata element, which are calculated for the words or kanji characters by using a training dataset and mapping rules. For instance, if the name of a metadata element has the character '名' (name), increase the suitability score for TITLE by 1, for PUBLISHER by 0.5, and for AUTHOR by 1. The character '名' (name) could exist in a word that has the semantics 'title', 'publisher', 'author', etc., and consists of combinations of several other kanji characters.

Our study of 334 metadata elements of 50 Japanese humanities digital libraries showed that 65 different elements have a potential to be regarded as TITLE, 46 as AUTHOR, 25 as SUBJECT, 77 as DESCRIPTION, 22 as PUBLISHER, 5 as TYPE, 20 as IDENTIFIER, 5 as SOURCE, 44 as COVERAGE, and 7 as RIGHTS. This shows how heterogeneous the metadata schemas of Japanese humanities digital libraries are, and it is vital to perform metadata mapping automatically.

According to the judgment of a native Japanese speaker who experienced in Japanese humanities digital databases and checked the results obtained when our automatic metadata mapping method mapped 334 attribute names of Japanese humanities collections to metadata elements of the modified DCMES, the average mapping precisions ranged from 85.7% to 100% (Table 2). Table 3 lists the average precisions we obtained using standard DCMES without or with the mapping rules and using modified DCMES with the mapping rules. The mapping precision obtained using modified DCMES with the mapping rules was 21.1 percentage points higher than that obtained using the standard DCMES without the mapping rules; this shows that the mapping rules improve the metadata mapping considerably. The average precision obtained using modified DCMES with the mapping rules was 15.9 percentage points higher than that obtained using the standard DCMES without mapping rules; this shows that the modified DCMES also improves the metadata mapping.

Table 2 Mapping precision of the automatic metadata mapping method

Modified DCMES element	Average precision (%)
TITLE	89.9
SUBJECT	100.0
AUTHOR	91.8
PUBLISHER	85.7
IDENTIFIER	100.0

DCMES: Dublin Core Metadata Element Set

Table 3 Comparison of metadata mapping precision

Metadata element set	Condition	Average precision (%)
Standard Dublin core	Without mapping rules	73.8
	With mapping rules	79.0
Modified Dublin core	With mapping rules	94.9

4.2 Retrieval in a federated searching system using automatic metadata mapping

To examine the performance of our federated searching system using automatic metadata mapping, we conducted an experiment by inputting a single query to three humanities collections (Ukiyo-e image database, donated Japanese books database, and old Japanese books database). Retrieval results obtained from these three collections for the sample query '風流' (elegance) in the TITLE metadata fields are shown in Fig. 3. Retrieval with other sample queries was also successful.

In addition to retrieving multiple humanities collections by inputting a single query in Japanese, we tested another function of our federated searching system in retrieving Japanese collections using an English query. This feature is very useful for users who do not understand Japanese; it allows searching and browsing Japanese digital libraries in English through a single interface and a single query (Batjargal *et al*., 2010). We applied this feature to the Ukiyo-e image database of the Art Research Center of Ritsumeikan University, which is freely accessible in Japanese.

As shown in Fig. 4, the Ukiyo-e artist name Kuniyoshi as an input query was translated as 国芳

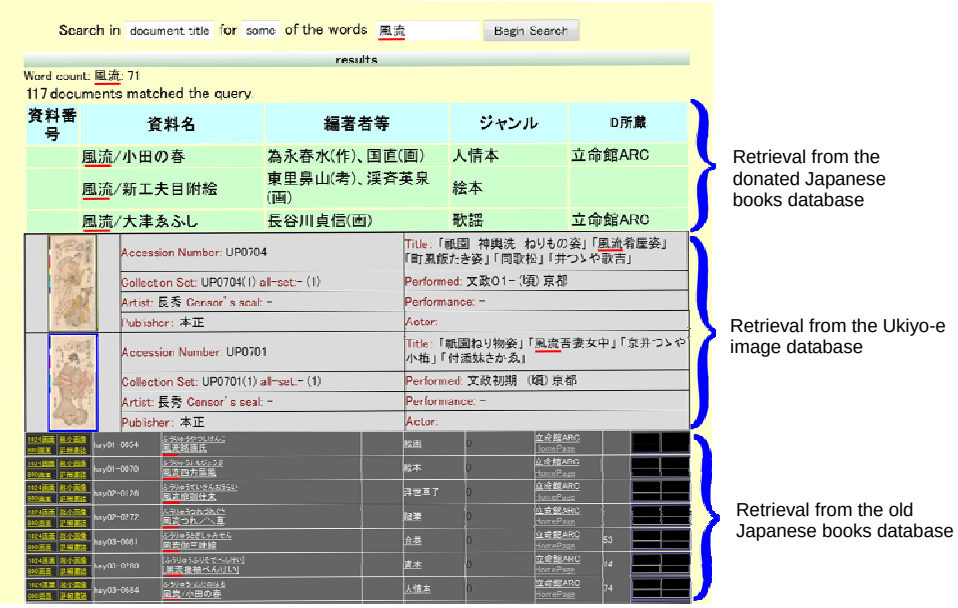

Fig. 3 Retrieval results obtained from three Japanese humanities digital libraries when using automatic metadata mapping

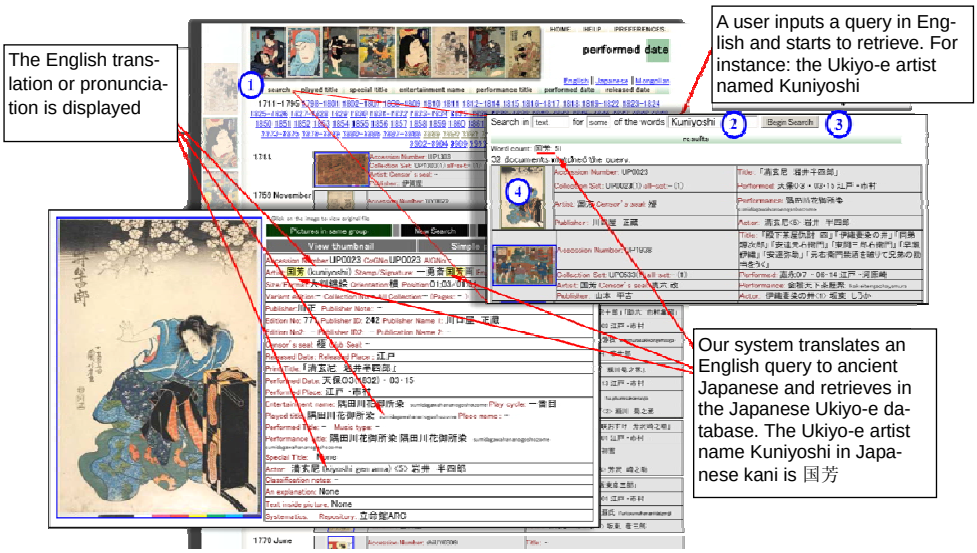

Fig. 4 Using an English query to search Japanese Ukiyo-e databases

and retrieved from the Japanese Ukiyo-e image database. The translated terms, names, explanations, etc. were displayed in English pages. Multiterm queries were treated as words: the artist's full name, UtagawaKuniyoshi, was treated as 歌川 (Utagawa) and 国芳 (Kuniyoshi) but not as 歌川国芳. As illustrated in Fig. 4, users will be able to enter a query in English (②) after clicking the Search button (①). The query Kuniyoshi is translated as 国芳 when the Begin Search button is clicked (③), and the translated query is retrieved from the Japanese Ukiyo-e image database. Lastly, the user will be able to access the webpage (④) that displays detailed information of a certain Ukiyo-e print, where the metadata in Japanese is translated and displayed in English.

In addition to mapping Japanese humanities metadata into the modified DCMES automatically, the rest of the metadata elements or some mapped elements were used without omission, because these are useful for some purposes. Among the metadata

elements of Japanese humanities digital libraries, some elements such as '西暦/Gregorian calendar/', '和暦/Japanese era calendar/', and '...名よみ/Reading of …' exist. The Japanese era calendar scheme is a common calendar scheme used in Japan which identifies a year by the combination of the Japanese era name and the year number within the era. For example, the year 2010 is 'Heisei 22'. The Japanese era names and the year numbers, months, and days as well as years in the Gregorian calendar are used together or separately in Japanese humanities metadata schemas. In our system these dates are converted into the DCMES encoding scheme for DATE and used for browsing by Japanese calendar in the Japanese pages and Gregorian calendar in the English pages, respectively. The elements such as '…名よみ/Reading of …' are used to provide the pronunciation of Japanese characters—kanji in hiragana. We used these elements for browsing metadata fields by alphabetical order or hiragana.

MeCab (http://mecab.sourceforge.net/) part-of-speech and the morphological analyzer for Japanese is used for word-segmentation and to convert kanji characters to hiragana or Latin alphabet. In general, Japanese humanities collections are processed in assistance with MeCab so that archaic and modern words are segmented properly. A bilingual dictionary with the humanities terms and special words was used for the Romanization, word-segmentation, user interface translation as well as the retrieval.

5 Discussions and future work

In this paper, we introduced our approach to constructing a federated searching system for humanities digital libraries, which used the automatic metadata mapping. Our proposed system aimed to provide a universal access to users in Japanese heterogeneous humanities digital libraries using only one query input, by mapping Japanese humanities metadata schemas to the modified DCMES metadata elements automatically. Moreover, we introduced one feature of our system in retrieving Japanese collections using an English query, which could help users who are looking for information written in Japanese (ancient or modern) that they may or may not understand.

The proposed method could be particularly important for large digital library collections, which have contents in various different languages including Japanese. Although the system presented in this paper is developed primarily for humanities researchers, this system might also be useful for ordinary users for finding the wanted information from multiple digital libraries using only one query input. This will further contribute to providing universal access to information and enhancing any digital library project by increasing the awareness of the digital collections in Japanese.

In future work, we will improve our system by linking more humanities digital libraries. We will further make our system work as a multilingual federated searching system, which retrieves heterogeneous humanities digital libraries in multiple languages via a single query input.

References

Akama, R., Okamoto, T., Maezaki, S., Takaaki, K., Kiyofumi, K., Ryoko, M., Oya, A., Mika, T., 2010. Image-Database and Studies for Japanese Arts and Cultures. Nakanishiya Shuppan, Kyoto, p.154-170.

Batjargal, B., Kimura, F., Maeda, A., 2010. Approach to Cross-Language Retrieval for Japanese Traditional Fine Art: Ukiyo-e Database. Proc. 14th European Conf. on Research and Advanced Technology for Digital Libraries, p.518-521. [doi:10.1007/978-3-642-15464-5_71]

Chan, L.M., Zeng, M.L., 2006. Metadata interoperability and standardization—a study of methodology, Part I: achieving interoperability at the schema level. *D-Lib Mag.*, **12**(6):4-6. [doi:10.1045/june2006-zeng]

Dublin Core Metadata Initiative, 2008. Dublin Core Metadata Element Set, Version 1.1. Available from http://dublincore.org/documents/dces/

Kawashima, M., Akama, R., Yano, K., Hachimura, K., Inaba, M., 2009. New Directions in Digital Humanities for Japanese Arts and Cultures. Nakanishiya Shuppan, Kyoto, p.133-154.

Kimura, F., Toba, T., Tezuka, T., Maeda, A., 2009. Federated Searching System for Humanities Databases Using Automatic Metadata Mapping. Proc. 9th Int. Conf. on Dublin Core and Metadata Applications, p.139-140.

Zeng, M.L., Jian, Q., 2008. Metadata. Neal-Schuman, New York, NY, p.3-5. [doi:10.1080/01639370902758378]