DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011

Metadata-related Challenges for Realizing a Federated Searching System for Japanese Humanities Databases

Biligsaikhan Batjargal
Ritsumeikan University, Japan
biligsaikhan@gmail.com

Fuminori Kimura
Ritsumeikan University, Japan
fkimura@is.ritsumei.ac.jp

Akira Maeda
Ritsumeikan University, Japan
amaeda@is.ritsumei.ac.jp

Abstract

This paper provides a summary of our ongoing project for providing integrated access to Japanese multiple digital libraries, archives, and museums. The main goal to construct a federated searching system for Japanese humanities databases, which searches multiple databases in parallel and provides on-the-fly integration of the results, has required the system to deal with heterogeneous metadata schemas in various formats. In this paper we discuss the metadata-related challenges faced at the front-end for retrieving multiple Japanese databases in parallel and integrating bilingual retrieved results. Aggregation and integration of the retrieved results in English and Japanese are complicated if a search needs to be performed from multilingual sources.

Keywords: federated searching; metadata mapping; Japanese humanities databases

1. Introduction

As a result of worldwide digitization over the last decade, many cultural institutions including libraries, archives, and museums started to expose digital objects on the Internet. Large digital library projects such as Europeana, World Digital Library, HathiTrust, and Google Book Search have been initiated and there has been progress in cross-domain collaboration as well as data exchange in recent years. Some institutions are collecting and archiving already available diverse Internet resources by taking advantages of metadata mapping, crosswalks and application profiles. However, metadata interoperability still remains a major issue. Different institutions use different ways to store metadata records of digital objects from their extensive collections. As a result, various digital objects are accessible in an inconsistent way through a variety of interfaces.

In this paper we address the metadata-related challenges faced when providing users access to multiple databases that have different metadata schema. We are developing a prototype federated searching system for Japanese traditional fine art: Ukiyo-e –woodblock printing databases, which retrieves multiple and heterogeneous back-end database servers including Z39.50, Search/Retrieve Web service (SRW)/Search/Retrieve via URL (SRU), OpenSearch, and other Web databases. We are developing our system by customizing and utilizing freely available open source software, such as Pazpar2, YAZ, SimpleServer, etc.

2. Japanese Writing Systems

Japanese text is written in a mixture of two writing systems—one using ideographic symbols, or *kanji*, and the other using *kana*, which consists of the syllabary scripts *hiragana* and *katakana*. A single kanji can have many pronunciations and can be used differently in words comprising two or more kanji. Therefore, it is helpful for users who don't know the Japanese language to know the pronunciation or reading (*yomi* in Japanese), which gives phonetic representation of a certain word. Usually, yomi is written in kana, although *romaji* (romanized representation of Japanese) is also used for transcriptions. There are several different romaji, but the revised Hepburn system, Kunrei-shiki Rōmaji (ISO 3602), and Nihon-shiki Rōmaji (ISO 3602 Strict) are

https://doi.org/10.23106/dcmi.952135776

widely used. Documents that have been catalogued in Japan are often transcribed according to the Kunrei-shiki Rōmaji or Nihon-shiki Rōmaji, whilst those cataloged in all the other countries of the world are transcribed according to the revised Hepburn system. Attaching the yomi or transcriptions in the metadata is important for Japanese databases because of the kanji's ambiguities for the given context. However, every database uses different solutions to store Japanese text and manage yomi, which results in some inconsistencies.

3. Federated Searching Approach for Japanese Content

In our prototype, search targets can be divided into two categories: 1) standards compliant that follow well-known standards for cataloguing, building metadata and exchanging data, and 2) web databases that have domain-specific metadata schemas which can be made accessible through different interfaces or protocols. Although records returned from multiple servers can be in different metadata formats or in HTML, we chose a simple schema based on Dublin Core Metadata Element Set (DCMES) at the front-end in order to integrate returned results on-the-fly. Basic elements such as <title>, <creator>, <subject>, <description>, etc. are adopted for displaying returned records as an integrated result listing in a single interface. In addition to the features of federated searching systems, our prototype has some features such as: (1) extracting yomi or transcriptions from Japanese text, (2) finding a translation if available and (3) displaying yomi (kana or romaji), translation, and kanji in a user friendly way. Some users may prefer original text in kanji, but others prefer translations. Some may understand Japanese kanji or kana, but romaji might be more readable for others. Yomi might be needed to know the correct pronunciation.

3.1. Standards Compliant Databases

This section discusses the challenges faced at the front-end for retrieving Japanese metadata records via standardized APIs, protocols and metadata schemas or search interfaces that are provided by various libraries and databases. In our prototype, remote databases can be searched and retrieved via the Z39.50, SRW/SRU, OpenSearch, etc. Results could be returned in any type of format that might include structured but plain text like SUTRS or data marked up in XML vocabularies such as Dublin Core, MARCXML, MODS, etc. Some important concerns for parsing Japanese text in the search results are explained here and example result sets obtained from various databases are shown in Figures 1-5. The metadata fields <title> along with the qualifier <transcriptions> are chosen to illustrate the challenges that are faced. Example result sets from the Library of Congress (LOC) matching the text query 'Hiroshige' in MARCXML and MODS are shown in Figure 1 and Figure 2. As shown in Figure 1, we are able to obtain Japanese text written in kanji, if we request a result set of return records in MARCXML. Requesting detailed records in MARCXML allows the display of original Japanese text, which cannot be obtained from the LOC in MODS (Figure 2). In MARC21 or MARCXML the 880 field – "Alternate Graphic Representation" replicates metadata in another field of the same record in a different script e.g., in kanji.

Another example as shown in Figure 3 is the result set from the British Museum retrieved via OpenSearch. The result provides very basic information and some information such as the name of print artist needs to be extracted from the <content>.

Furthermore, Figure 4 shows some elements in the result set from the National Diet Library (NDL) of Japan in the NDL Metadata Element Set, which is standardized by adding qualifiers to the DCMES. Some useful data that we were able to get in Figure 4 are English translation of the <title> and <transcriptions> in kana, which cannot be obtained in Dublin Core. However, more detailed information perhaps in JAPAN/MARC, or METS based NDL Digital Archiving System Metadata Schemas cannot be obtained through the NDL APIs.

The field <dc:title> is useful for displaying English translation of the title at the back end. However, it might be not as easy as just extracting values, because we might get transcriptions

https://doi.org/10.23106/dcmi.952135776

instead of translation (Figure 5), which is quite different from <dcndl:titleTranscription>. It requires performing some additional tasks such as language detection, etc.

```
<datafield tag="245" ind1="0" ind2="0">
  <subfield code="6">880-01</subfield>
  <subfield code="a">Kyōka hyakunin isshu /
</subfield>
  <subfield code="c">Tenmei Rōjin Takumi kō ;
Ryūsai Hiroshige ga.</subfield></datafield>
<datafield tag="630" ind1="0" ind2="0">
  <subfield code="6">880-03</subfield>
  <subfield code="a">Ogura hyakunin
isshu</subfield>
  <subfield code="v">Parodies, imitations,
</subfield></datafield>
<datafield tag="700" ind1="1" ind2=" ">
  <subfield code="6">880-05</subfield>
  <subfield code="a">Andō, Hiroshige,</subfield>
  <subfield code="d">1797-
1858.</subfield></datafield>
<datafield tag="880" ind1="0" ind2="0">
  <subfield code="6">245-01/$1</subfield>
  <subfield code="a">狂歌百人一首 /</subfield>
  <subfield code="c">天明老人内匠校 ； 立齋廣重画
.</subfield></datafield>
<datafield tag="880" ind1="0" ind2="4">
  <subfield code="6">630-03/$1</subfield>
  <subfield code="a">小倉百人一首</subfield>
  <subfield code="v">Parodies, imitations,
etc.</subfield></datafield>
<datafield tag="880" ind1="1" ind2=" ">
  <subfield code="6">700-05/$1</subfield>
  <subfield code="a">安藤廣重,</subfield>
 <subfield code="d">1797-1858.</subfield></datafield>
```

FIG. 1. Sample result set of the LOC in MARCXML.

```
<titleInfo>
<title>Kyōka hyakunin
isshu</title>
</titleInfo>
<name
type="personal">
<namePart>Tenmei
Rōjin</namePart>
<namePart
type="date">1781-
1861</namePart>
</name>
<name
type="personal">
<namePart>Andō,
Hiroshige</namePart>
<namePart
type="date">1797-
1858</namePart>
</name>
<subject
authority="lcsh">
 <titleInfo>
 <title>Ogura hyakunin
isshu</title>
</titleInfo>
<genre>Parodies,
imitations, etc</genre>
</subject>
```

FIG. 2. Sample result set of the LOC in MODS.

```
<entry><title>Kyoka
Hyakunin Isshu 狂歌百
人一首</title>
<link
href="http://www.british
museum.org/research/se
arch_the_collection_dat
abase/search_object_de
tails.aspx?objectid=7792
22&amp;partid=1">
</link>
 <content
type="text">illustrated
book ; Utagawa
Hiroshige (歌川広重)
(Made by);  Illustrated
book. 1 vol. Woodblock-
printed</content>
<id>urn:uuid:e5d2dca4-
6f0a-4b57-b501-
ce44088383f1</id>
<updated>2011-07-
05T11:28:02+01:00
</updated>
<nmlop:imageurl></nmlo
p:imageurl>
<relevance:score>0.82
</relevance:score>
</entry>
```

FIG. 3. Sample result set from the British Museum.

```
<dc:title>山水面白く、また物凄し--広重日記に見る情緒性 =The
view was enjoyable and also ghastly: affection expressed in
Hiroshige's Diary ( 特集　日本美術の叙情性--情趣の系譜
)</dc:title>
<dcndl:titleTranscription>サンスイ　オモシロク　マタ　モノスゴシ
ヒロシゲ　ニッキ　ニ　ミル　ジョウチョセイ　( トクシュウ　ニホン
ビジュツ　ノ　ジョジョウセイ　ジョウシュ　ノ　ケイフ
) </dcndl:titleTranscription>
<dc:creator>岸　文和</dc:creator>
<dcndl:creatorTranscription>キシ　フミカズ
</dcndl:creatorTranscription>
<dc:subject>歌川 広重 </dc:subject>
<dcndl:subjectTranscription>ウタガワ　ヒロシゲ
</dcndl:subjectTranscription>
```

FIG. 4. Sample result set of the NDL in the NDL Metadata Element Set.

```
<dc:title>太田牛一『豊国大明神臨時御
祭礼記録』の諸本と改稿の意味 =Ota
Gyuichi "Toyokuni-daimyojin rinjigosairei
kiroku"</dc:title>
<dcndl:transcription>オオタギュウ　イチ
トヨクニ　ダイ　ミョウジン　リンジ　ゴ
サイレイ　キロク　ノ　ショホン　ト　カイ
コウ　ノ　イミ</dcndl:transcription>
<dcndl:transcription>OOTAGYUU ICHI
TOYOKUNI DAI MYOUJIN RINJI GO
SAIREI KIROKU NO SHOHON TO
KAIKOU NO IMI</dcndl:transcription>
```

FIG. 5. Odd result set of the NDL in the NDL Metadata Element Set.

3.2. Non-standard databases

In this section we discuss the challenges for integrating the retrieved results of Japanese Web databases that do not support well-known standards for cataloguing, building metadata and exchanging data. Here we utilized "screen-scraping" techniques that process a list of search

https://doi.org/10.23106/dcmi.952135776

results by reading and extracting the HTML. Many Japanese databases developed metadata based on their domain-specific semantics and content. Names or labels for metadata elements are written in Japanese, or labels in Japanese are used as the metadata elements. It is necessary to understand the semantics of Japanese databases–such as elements, syntax, and structure– in order to perform metadata mapping to a well-known schema, especially when it is written without explicit word boundaries and uses short forms. The absence of word delimiters makes word segmentation 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 extracting metadata elements, indexing and metadata mapping. The metadata element names used in Japanese databases consist of several words that have combinations of single or several kanji characters, and the meaning of the words depends on the combinations. Therefore, we utilized the automatic metadata mapping method (Batjargal et al., 2010; Kimura et al., 2009) for identifying titles, authors, and other fields in Japanese text. For instance, for the Ukiyo-e image database of the Ritsumeikan University's Art Research Center, 10 out of 67 field names such as "画題等", "画題２", "役名", "外題", "所作題", "細目題", "主外題", "系統分類題", "演目(統合)", and "画題統合" that were written in various kanji characters were identified and mapped to <title>, though all have different meanings, such as "Print title", "Picture name", "Official title", "Title of play", "Performed title", "Detailed title", "Main performed title", "Classification title", "Name of musical", "Title of the integrated picture", and "Material name" respectively. Similarly, 6 elements such as "外題よみ", "所作題よみ", "細目題よみ", "主外題よみ", "演目よみ", "資料名よみ" were mapped to <transcriptions>. In this case, yomi or reading was written in hiragana, and was different from NDL's <dcndl:titleTranscription> that was written in katakana. The elements <title> and <transcriptions> both may have multiple values. Original values in Japanese kanji need to be included as multiple value strings.

After gathering search results and mapping to simple elements including <title>, <creator>, <subject>, <description>, etc along with their <transcriptions>, our system will parse and aggregate the search results and display them in a user-friendly way. At this point, some discussion about parsing the results is required, since metadata values or data may be returned in an inconsistent format between databases. For example, full-width (zenkaku) alphanumeric characters, i.e., two-byte characters, are widely used in Japanese text. In order to perform further tasks such as merging, relevance ranking, sorting, and showing faceted results, all full-width alphanumeric characters should be converted to ASCII. Another consideration is that, a date may appear as "April 30, 2011" in one record, as "1/4/2011" in another and as "平成 23 年 4 月 30 日" in Japanese. 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 2011 is "平成 23/Heisei 23/". The format of a date field has to be normalized across all results from all sources. However, the Japanese calendar could be displayed in the Japanese pages.

The <transcriptions> in romaji, <title> in English, and the translation of Japanese <title> could be used for displaying Japanese content in English pages. Meanwhile, the <transcriptions> in kana, <title> in Japanese, and the translation of English <title> could be displayed in Japanese pages. Such a feature is very useful for users who do not understand Japanese, and it allows searching and browsing multiple Japanese digital libraries in parallel (Batjargal et al., 2010a). In this way, users are provided with a list of bilingual documents, holding basic information and a link where the original documents can be found. Figure 8 shows the screenshot of our federated searching system, which in the earlier stage of development.

4. Summary and Findings

In this paper we discussed the metadata-related challenges faced at the front-end for retrieving multiple Japanese databases in parallel and integrating retrieved results on-the-fly. We accessed heterogeneous databases in English and Japanese that share data in various formats; simple or

https://doi.org/10.23106/dcmi.952135776

complex, such as Dublin Core, MARCXML, MODS, etc. We acknowledged that a better way for displaying returned records as an integrated result listing in a single interface is to select basic elements, e.g., <title>, <creator>, <subject>, <description>, <date> for metadata mapping. However, aggregation or integration of the retrieved results could be complicated if a search needs to be performed from multilingual sources. Furthermore, obtaining records in as much detail as possible helps to avoid losing richness and meaning of native metadata. Moreover, attaching yomi or transcriptions that denote the pronunciation in the metadata is important for Japanese databases because of the kanji's ambiguities for the given context. We recognized that language tags for the metadata value string are not enough to distinguish scripts, since parallel written yomi or transcriptions of Japanese text could be written in hiragana, katakana, or romaji.

For the future work, collaborative searching, harvesting or indexing techniques can be adopted, in order to improve the proposed system allowing quick searching and effective ranked results.

FIG. 8. Returned records as an integrated result list in a single interface

References

Akira Miyazawa. (2007). Parallel writing in East Asian languages and its representation in metadata in light of the DCMI Abstract Model. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2007, 1-9.

Batjargal, Biligsaikhan, Fuminori Kimura, and Akira Maeda. (2010). Providing Universal Access to Japanese Humanities Databases: An Approach to Federated Searching System Using Automatic Metadata Mapping. Journal of Zhejiang University-SCIENCE C, Vol. 11, No. 11, 2010, 837-843

Batjargal, Biligsaikhan, Fuminori Kimura, and Akira Maeda. (2010a). Approach to Cross-Language Retrieval for Japanese Traditional Fine Art: Ukiyo-e Database. Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2010), 2010, 518-521

https://doi.org/10.23106/dcmi.952135776

DCMI. (1999). Dublin Core Metadata Element Set, version 1.1: Reference description. Retrieved April 25, 2011, from http://dublincore.org/documents/1999/07/02/dces/.

DCMI. (2010). DCMI Metadata Terms. Retrieved April 25, 2011, from http://dublincore.org/documents/dcmi-terms/.

Kimura, Fuminori, Takushi Toba, Taro Tezuka, and Akira Maeda. (2009). Federated Searching System for Humanities Databases Using Automatic Metadata Mapping. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2009, 139-140.

Library of Congress. (2010). MARCXML. Retrieved April 25, 2011, from http://www.loc.gov/standards/marcxml/.

https://doi.org/10.23106/dcmi.952135776