2

Integrated Information Access Technology for Digital Libraries: Access across Languages, Periods, and Cultures

Biligsaikhan Batjargal,[1] Garmaabazar Khaltarkhuu,[2]
Fuminori Kimura[1] and Akira Maeda[1]
[1]*Ritsumeikan University*
[2]*Mongolia-Japan Center for Human Resources Development*
[1]*Japan*
[2]*Mongolia*

1. Introduction

Physical libraries store materials written in various languages, at various periods in history, and dealing with various cultures. As a result, large digital library projects such as Europeana, World Digital Library, HathiTrust, and Google Book Search have collections spanning different languages, periods, and cultures. This diversity complicates information access, in part because the grammars, vocabularies, and scripts of languages usually change significantly over time.

This chapter presents our approach to providing cross-language access that accounts for this evolution of languages over periods ranging from ancient to modern and even considers cultural differences. It also presents our method for providing integrated access to multiple digital libraries, archives, and museums by automatically mapping between different metadata schemas. In section 2, we present the traditional Mongolian script digital library. Our proposed method for Cross-period information retrieval from ancient Japanese historical Materials is discussed in section 3. Later, in section 4, we introduce the federated searching system for humanities databases using automatic metadata mapping.

2. Traditional Mongolian script digital library

In recent years the importance of digital cultural heritage preservation has been increasing in the Asia-Pacific region as well as worldwide. This section provides a summary of the recent achievements of the Traditional Mongolian Script Digital Library (TMSDL) (Khaltarkhuu et al, 2007; Khaltarkhuu et al, 2008), which aims to preserve over 800 years of historical records written in traditional Mongolian for future use and to make them available for public viewing. There are over 50,000 registered manuscripts and historical records written in traditional Mongolian script stored in the National Library of Mongolia. About 21,100 of them are handwritten documents and over 9400 of those are related to the history of Mongolia (Tungalag, 2005). Despite the importance of keeping old historical materials in good conditions, the Mongolian environment for material storage is not suitable

for keeping historical records for a long time (Tungalag, 2005). An efficient and effective way to preserve and protect materials of historical importance while making them publicly available is to digitize them and create a digital library.
Mongolian is spoken by most of the Mongolian population as well as by Inner Mongolians and other groups of people who live in several provinces of China and the Russian Federation. It is one of the many languages of the Mongol-Altaic family.
Mongolians have used numerous writing systems, and traditional Mongolian script is the longest surviving script. Although the Mongolian language has also been written in Chinese characters, Phags-pa script, Soyombo script, Horizontal square script, Latin, and Cyrillic script (Shagdarsuren, 2001), by the end of the 20th century the traditional Mongolian script had made an official, government-decreed return. At the result, modern Mongolian language has two distinct writing systems: Cyrillic and traditional Mongolian.
The sounds of words changed as the Mongolian language evolved, but the spelling remained unchanged. Thus was created a difference between written Mongolian and spoken Mongolian. However, in 1946 in Mongolia the Cyrillic script was adopted with two additional characters. At that time the spelling of modern Mongolian in the Cyrillic alphabet was based on the pronunciation of the dialect spoken by the Khalkha, a subgroup of the Mongols. This was a radical change and alienated the Mongolian people from their culture and historical archives written in traditional Mongolian script. Traditional Mongolian script preserves a more ancient language and reflects the Mongolian language spoken in the ancient period, while modern Mongolian reflects pronunciation differences in modern dialects. Traditional Mongolian is a distinct dialect with grammar different from that of modern Mongolian. The traditional Mongolian script is written vertically, from top to bottom, in columns advancing from left to right. This script is the writing system for the Mongolian language and has four derivative scripts: Todo, Manchu, Vaghintara, and Sibe (Xibe). The Todo script was used by the Oirats and Kalmyks, and the Manchu script was a writing system in the Qing dynasty. The Sibe script is used in Xinjiang, in the northwest of China. The Vaghintara script was used by the Buryats. Like Arabic, traditional Mongolian is a contextual script where letters are cursively joined and have initial, medial, and final presentation forms for the same letter. In most cases the letters join together along a vertical stem, but in the case of certain consonants that lack a trailing vertical stem they may form a single ligature with a following vowel. In addition to these cursive and positional forms, many letters also have variant forms used in accordance with spelling and grammatical rules.
Using modern Mongolian to retrieve information from traditional Mongolian documents is not a simple task because the Mongolian language has changed substantially over time. The traditional Mongolian script digital library (TMSDL) (Khaltarkhuu et al, 2007; Khaltarkhuu et al, 2008), which is based on Greenstone Digital Library Software (GSDL) and accepts modern Mongolian query input, will help the user access materials written in traditional Mongolian.

2.1 Ancient-to-modern information retrieval

Thanks to advances in innovative information technologies and to the popularity of the Internet, many ancient historical documents are being digitized and made publicly available. We therefore want to offer an "ancient-to-modern information retrieval" method (Batjargal et al., 2010a; Batjargal et al., 2010b) that considers language differences over time. We aimed to develop a retrieval system with which a user can access cross-period and cross-script ancient document databases by using a query in a modern language.

There has been little research on information retrieval techniques for historical documents, and almost none of the breakthroughs in research on information retrieval and information access have aimed at retrieving information in the native language from ancient, cross-period and/or cross-script foreign language documents.

Few approaches that could be considered a cross-period information retrieval have been proposed (Ernst-Gerlach & Fuhr, 2007; Koolen et al., 2006; Gotscharek et al., 2009; Hauser et al.,2007; Pilz et al., 2008), and there has been little research on information retrieval techniques for historical documents. (Ernst-Gerlach & Fuhr, 2007) focused on modern and archaic German and developed a retrieval method that considers the spelling differences and variations over time. (Koolen et al., 2006) considered the spelling and pronunciation differences between ancient and modern Dutch, while (Gotscharek et al., 2009) and (Hauser et al., 2007) considered the spelling differences and variations between modern and archaic German. (Pilz et al., 2008) considered spelling variations of English and German historical texts. In general, the main challenge for historical European languages like Dutch, English, and German is the spelling variants.

We applied an "ancient-to-modern information retrieval" method to ancient Mongolian historical collections written in traditional Mongolian script. Some ancient historical documents in traditional Mongolian script have recently been digitized and made publicly available, and text-display support for traditional Mongolian script and the input locale is enabled in Windows Vista and Windows 7. The Uniscribe–Unicode Scripts Processor driver was updated to support OpenType advanced typographic functionality of complex text layouts, such as traditional Mongolian script.

The situation for an ancient Mongolian language is a bit different because the Mongols have changed their writing systems several times and more than once have made language reforms that eliminate a difference between written and spoken language (Shagdarsuren, 2001).

2.2 Proposed approach

To cope with cross-period and cross-script Mongolian documents, we propose a simple model that retrieves traditional Mongolian documents using modern Mongolian query. The structure of the TMSDL (Khaltarkhuu et al, 2007; Khaltarkhuu et al, 2008), with the proposed "ancient-to-modern information retrieval" approach (Batjargal et al., 2010a; Batjargal et al., 2010b) is shown in Fig. 1. We utilized the existing approach (Kimura et al., 2009) and improved the "retrieval technique with the modern Mongolian query on traditional Mongolian text" (Khaltarkhuu et al, 2006) by integrating a dictionary. A prototype of the TMSDL (Batjargal et al., 2010a; Batjargal et al., 2010b), which could be considered a cross-period information retrieval system, has been developed. The retrieval method of the TMSDL considers cross-period differences in the writing systems of the ancient and modern Mongolian languages. Adding a dictionary-based query translation approach to the translation module was a major improvement that takes into account age differences in the writing systems of the ancient and modern Mongolian languages. We utilized the developing online version of Tsevel's concise Mongolian dictionary (Tsevel, 1966) under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported license. Tsevel's dictionary was printed in 1966 and is one of two Mongolian dictionaries with definitions written in modern and traditional Mongolian available on the market. It includes over 30,000 words in Cyrillic and traditional Mongolian script.

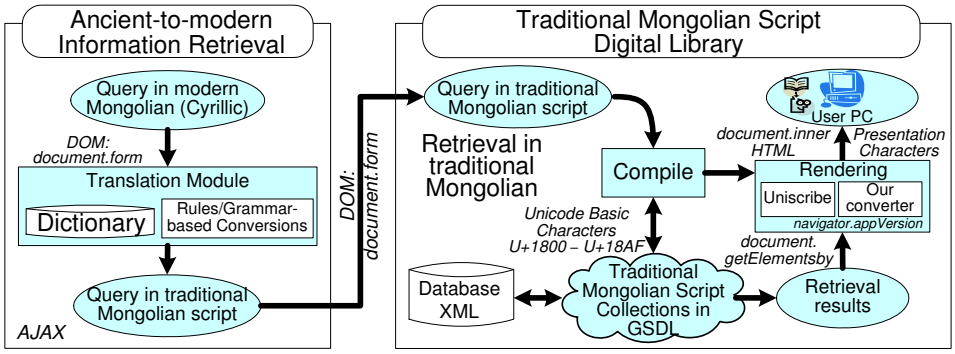


Fig. 1. Ancient-to-modern information retrieval in the TMSDL.

To boost the quality of the translation, the "ancient-to-modern information retrieval" approach (Batjargal et al., 2010a; Batjargal et al., 2010b) matches query terms to words in Tsevel's dictionary. If no exact match is found, the "retrieval technique with the modern Mongolian query on traditional Mongolian text" (Khaltarkhuu et al, 2006), which is based on grammatical rules, is used. The proposed model allows the users to access documents written in an ancient language (traditional Mongolian) with a query input in a modern language (modern Mongolian – Cyrillic). As shown in Fig. 1, the query in modern Mongolian (Cyrillic) is translated into a query in traditional Mongolian script. The query in traditional Mongolian (Unicode characters in the range U+1800 – U+18AF) is then submitted as a retrieval query for traditional Mongolian script collections. Chronological books of ancient Mongolian kings, Genghis Khan, and the Mongol Empire (the largest contiguous empire in history) such as the Altan Tobci (year 1604, 164 pp) and the Story of Asragch (year 1677, 130 pp) etc, are available in the TMSDL with a modern Mongolian input interface. A database of such historical records with a modern language query input will help someone conducting research on the history of the High Middle Ages understand 13th–14th century history of Asia. The modern Mongolian (Cyrillic) input in the TMSDL is illustrated in Fig. 2.

2.3 Experimental evaluation

In an experiment we conducted in order to check the correctness of translations from the modern language to the ancient one, we retrieved traditional Mongolian documents when using modern Mongolian query input in Cyrillic. Because of the large number of unfamiliar ancient proper nouns, terms, and their variants in ancient historical documents, we faced the challenge of measuring recall and precision as well as the challenge of defining relevant documents. To check whether a queries in modern Mongolian (Cyrillic) were translated correctly, we selected queries the most frequently appearing words that are pronounced or written differently in modern and traditional Mongolian and compared their word counts in the search results with the corresponding word counts in "Qad-un ündüsün quriyangγui altan tobči –Textological Study" (Choimaa & Shagdarsuren, 2002). This textological study contains a detailed analysis of traditional Mongolian word frequencies in the Altan Tobci.

We compared the word count in the search results for two cases: one using only grammatical-rule-based translation, and the other additionally using a dictionary. The version with dictionary integration translated and retrieved 86% of the input queries, whereas the grammatical-rule-based version retrieved only 61% of the input queries. Even

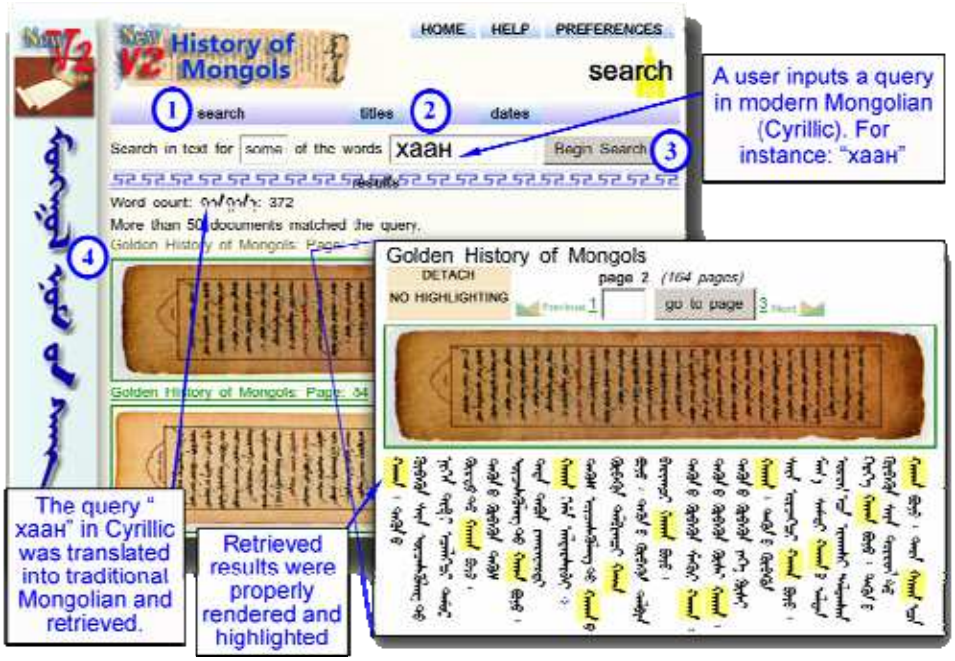


Fig. 2. TMSDL Cyrillic input and retrieval results.

with the dictionary, however, 64% of the input queries in modern Mongolian did not match with a word count that was less than or greater than the actual number (frequency) because of possible errors of translation, grammatical inflection, and text digitization, or limitations of the indexer and retrieval function. Comparisons of the retrieval results are illustrated in Fig. 3, and detailed retrieval results for sample query terms are shown in Table 1 along with modern and ancient forms, their meanings, and the word counts. A retrieval result with the query word highlighted is shown in Fig. 2.

The TMSDL integrated with a dictionary translated and retrieved 86% of the input queries, but only 22% were retrieved without error.

2.4 Summary and future directions

In this section we introduced the TMSDL that utilizes cross-period and cross-script digital collections and that enables historical documents written in an ancient language to be accessed using a query in a modern language. The proposed system is suitable for full text searches on databases containing cross-period and cross-script documents. Such research would involve extensive research in an ancient language that users and humanities researchers may or may not understand. It could apply to humanities researchers who are conducting research on ancient culture and looking for relevant historical materials written in that ancient language. The proposed model will enable users and humanities researchers to search for such materials easily in a modern language. We still, however, need improvements dealing with such problems as a total failure to translate 14% of input queries. Improvements in translation and retrieval techniques also need to be considered.

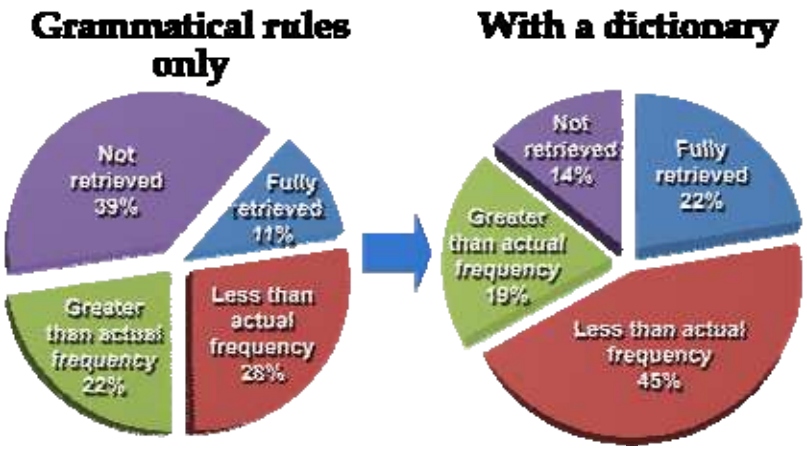


Fig. 3. TMSDL retrieval results obtained using different translation methods.

Word in Modern Mongolian (Cyrillic)	Pronunciation		Word in traditional Mongolian	English translation	Word count in retrieval results		Actual word count	Retrieval state
	Modern	Ancient			Grammatical rules only	With a dictionary		
хүн	hŭn	kŭmŭn		man, person, human	0	135	135	Fully retrieved
хаан	qaan	qaγan		king	0	372	372	
даян	dayan	dayan		all, whole	0	0	32	Not retrieved
сайн	sain	sayin		good, well, fine, nice, pretty	0	0	75	
гэр	ger	ger		home, residence	0	51	47	Greater than actual frequency
төр	tör	törü		law, kingdom	0	50	47	
эзэн	ezen	eǰen		lord	0	144	146	Less than actual frequency
богд	bogd	boγda		holy, sacred, divine	0	39	40	

Table 1. Examples of retrieval performance obtained using different translation methods.

Two interesting subjects for future research are the retrieval of information from two distinct ancient languages and using a single query input in a modern language for retrieval from multiple sources in multiple ancient languages.

The next section discusses the our another achievement – cross-period information retrieval method for ancient Japanese historical materials

3. Cross-period information retrieval from ancient Japanese historical materials

Libraries, governments, and major internet providers have recently begun forming consortiums to preserve historical documents stored in libraries. This means that more and more old-text content will soon be accessible on the Internet. The huge amount of knowledge in old documents is obviously as important as that in the recently created digital documents typically available on the web because old documents contain the wisdom of our ancestors.
Retrieving important information from old documents is not always easy, however, because languages and cultures change substantially over time. To access documents written in ancient Japanese by using a query in modern Japanese, for example, we need a cross-period information retrieval system based on a cross-period (ancient-modern) Japanese dictionary.

3.1 Construction of ancient-modern dictionary

Ancient documents in text form are being digitized, and the prevalence of search engines has made the retrieval of information from digital documents a familiar procedure. Current search engines, however, may be not able to acquire proper retrieval results for ancient Japanese documents because there is no ancient-modern Japanese dictionary with sufficient entries.
One reason for this is that the Japanese writing system has no term separation. That is, neither current nor ancient Japanese writing uses space or punctuation to separate words. A morphological analyzer like ChaSen or MeCab, both of which need a modern term dictionary, is usually used to do term separation for modern Japanese, but there is no ancient-modern word dictionary with enough entries and there are no morphological analyzers for ancient Japanese. This makes it difficult to do term separation for ancient Japanese.
We propose a method for constructing an ancient-modern Japanese dictionary by using a parallel corpus of ancient writings and their translations in modern Japanese. The parallel corpus thus consists of pairs of documents in the same language but in ancient and modern versions of that language. From this corpus we try to acquire pairs of equivalent archaic and modern words by analyzing the frequencies of word occurrences in a sentence in ancient Japanese and its corresponding modern Japanese translation.

3.1.1 Related work

Two methods for extracting pairs of equivalent words from a bilingual corpus in modern languages (English and Japanese, for example) have already been proposed, one using a parallel corpus and the other using a non-parallel corpus. In the method using a parallel corpus, equivalence is based on statistical correlation determined using co-occurrence frequency, contingency tables, etc. (Kitamura & Matsumoto, 1996). In the method using a non-parallel corpus, equivalence is based on the context similarity of translation candidates (Tanaka, 2002). The method described here, however, identifies pairs of equivalent words not in two modern languages but in modern and archaic Japanese. As there are few modern language translations of ancient writings, it is difficult to collect a parallel corpus of ancient writings and their translations in modern language. Some famous ancient writings, though, have been translated into the modern forms of their languages. We therefore identify pairs of equivalent words in modern and archaic Japanese by using a parallel corpus comprising famous ancient Japanese writings and their translations in modern Japanese.

3.1.2 Proposed method of dictionary construction

Many well-known ancient writings have modern-language translations, and some of these translations are digitized and open to the public. In a parallel corpus comprising writings in an ancient and modern language, one can usually determine which modern-language sentence corresponds to which ancient-language sentence. A modern word equivalent to an archaic word in an ancient-language sentence is likely to appear in the modern-language translation of that sentence, and vice versa. Word pairs with high co-occurrence frequency in ancient and modern sentence pairs are thus likely to be translation equivalents.
In our method we detect similarities in the appearance tendencies of modern and archaic words in each sentence pair and then use these similarities to extract equivalent pairs of ancient and modern words (Fig. 4).

A. Word Extraction from Parallel Corpus

We use morphological analysis to extract words from the modern-language translations of the ancient writings, and because there is no morphological analyzer for ancient Japanese. We divide the archaic sentences into N-grams and treat those N-grams as archaic words.
An N-gram is a sequence of N characters from a given string. We first extract the first N characters from the target string and then shift one character and extract N characters from the target string. We repeat this shifting-and-extracting process until the N^{th} character in the N-gram is the last character of the target string. For example, the string "corpus" would be divided into the following four 3-grams: cor, orp, rpu, and pus.
One of the drawbacks of the N-gram approach is that there will be many overlaps. On the other hand, an advantage of the N-gram approach is that it can divide the strings even if the language of the string, like ancient Japanese, does not have explicit delimiters between words. This is why we divide the archaic sentences into N-grams and treat those N-grams as words.

B. Calculation of the Co-occurence of Modern and Archaic Words

In this process, we calculate co-occurrence frequencies of archaic terms and modern terms that are extracted in section 3.1.2.A. This process is conducted for archaic and modern term pairs to appear in the equivalent sentences. In other words, the term pairs appearing in the equivalent sentences are considered as the co-occurring terms.
In each sentence pair, the archaic and modern term pairs are created for every possible pairs of extracted modern terms and archaic N-grams. We count the occurrence frequency of each term pairs. This frequency is the co-occurrence frequency of archaic and modern term pairs.

C. Calculation of Similarity about Appearance of Tendency between Modern Term and Archaic Term

For parallel corpus composed two different languages documents such as Japanese and English, "mutual information" is proposed to use for the similarity between each two terms (Kitamura & Matsumoto, 1996). Our method also adopts "mutual information" in order to calculate similarities about appearance of tendency between modern term and archaic term.
The archaic and modern term pairs that have higher value of their mutual information is considered that appearance of tendency between modern term and archaic term is similar. These term pairs have higher possibility that the modern term is relation in translation for the archaic term. We extract term pairs that have higher similarities than some threshold, and consider that these pairs have relation in translation.

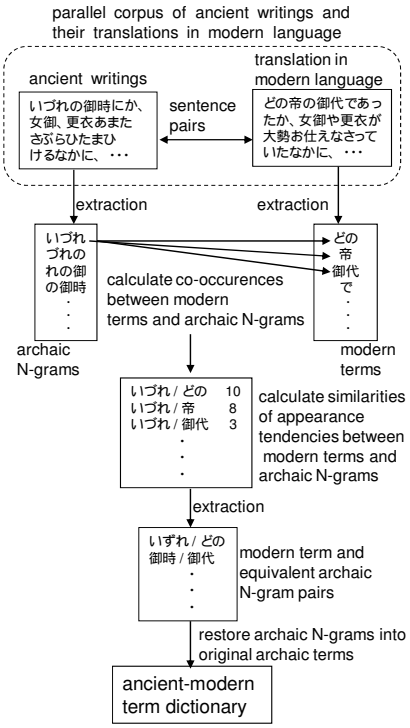


Fig. 4. Flow of the construction of an ancient-modern dictionary.

The mutual information MI of a modern term t and an archaic N-gram g is given by the following formula.

$$\mathrm{MI}(t,g) = \log \frac{P(t,g)}{P(t)P(g)} \tag{1}$$

Probability P(t, g) is the probability that the modern term t appears in the translation of the archaic sentence in which the archaic N-gram g appears, and it can be calculated from the co-occurrence frequency of archaic N-gram g and modern term t. Probabilities P(t) means the probability in the case that the modern term t appears in modern sentence. Probabilities P(g) means the probability in the case that the archaic N-gram g appears in archaic sentence. Probabilities P(g) is able to be acquired from the term frequency of archaic N-gram g as mentioned in section 3.1.2.A.

The archaic and modern term pairs that have higher value of their mutual information are considered that appearance of tendency between modern term and archaic term is similar.

D. Extraction of Translation Pairs of Modern and Archaic Words

In section 3.1.2.C, we extract archaic and modern term pairs that have higher possibilities of relation in translation. However, as the archaic terms of extracted pairs are represented by N-gram, these archaic terms are not always complete archaic term. Some archaic N-gram may be part of archaic term. Another may be combined parts of some archaic terms. In these

cases, we have to restore the archaic N-grams to original archaic terms. These archaic N-grams are restored to original archaic terms by comparing spellings, term frequency and co-occurrence frequency between another archaic N-gram. We consider that restored archaic and modern term pairs are related in translation. Finally, we collect these term pairs and construct ancient-modern term dictionary.

3.1.2 Future directions

We proposed a method for constructing an ancient-modern Japanese dictionary by using a parallel corpus of ancient writings and their translations in modern Japanese. If an ancient-modern Japanese dictionary with sufficient entries is constructed by the proposed method, we think that the techniques of natural language processing, for example morphological analysis, could be applied for ancient documents digitized in text form.

We need to improve the term extracting process in order to reduce the number of unnecessary word pairs, to improve the calculation of similarities of the appearance tendencies of modern and archaic words, and to construct a practical ancient-modern Japanese dictionary.

3.2 Cross-period information retrieval system

There has been a lot of research on cross-language information retrieval in the last decade. Various approaches—including query translation, document translation, and the use of an intermediate language—has been studied, and adequate retrieval effectiveness has been achieved for some pairs of languages (e.g., certain European languages).

There has, in contrast, been very little research on information retrieval methods for historical documents, and most of those methods are based on simple keyword matching. Some recently proposed approaches to accessing historical documents consider the evolution of languages and could be regarded as a kind of cross-age information retrieval (Gerlach & Fuhr, 2007; Khaltarkhuu & Maeda, 2006). Our goal is to establish a more effective and sophisticated retrieval method that considers not only language difference over time but also cultural differences between languages and ages.

The architecture or the cross-period information retrieval system we developed is shown in Fig. 5. This system lets old Japanese documents be retrieved using modern Japanese keywords, so old Japanese documents by users who do not know archaic Japanese.

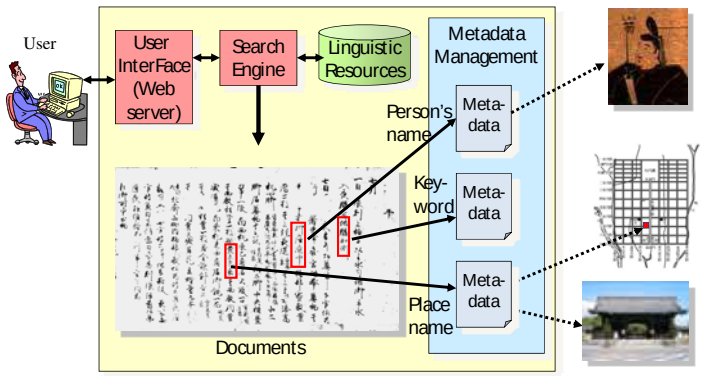


Fig. 5. Architecture of proposed cross-period information retrieval system for ancient Japanese historical materials.

3.2.1 Proposed method for cross-period information retrieval

We use the dictionary-based query translation approach because it is the one most effective for cross-language information retrieval. For dictionary-based methods to be effective we need to use precise and comprehensive dictionaries for both the modern and ancient language. We try to find relations between the entries in those dictionaries and to translate the query terms in the modern language into equivalent terms in the ancient language. For this translation process we propose the following method (Fig. 6).

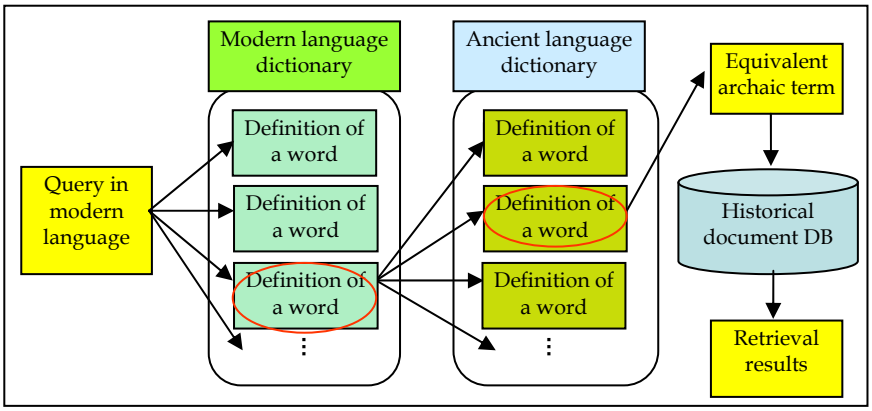


Fig. 6. Overview of proposed method for cross-period information retrieval.

1. For each entry in the modern-language dictionary, we look for an equivalent entry in the ancient language dictionary by calculating the similarities between the definition of the modern word and all the definitions of the archaic words. We can do this using a standard text similarity measure based on the vector space model and the tf-idf term weighting scheme.
2. We then take the most similar definition in the ancient language dictionary and regard the dictionary entry (headword) containing that definition as an equivalent of the modern word.
3. If there is more than one equivalent entry, we find the one most nearly equivalent to the modern word by using a term association measure such as mutual information to disambiguate the candidate translations.

3.2.2 Implementation

We implemented a cross-period information retrieval system for the Japanese historical document called the *Hyohanki* diary. Written in late Heian era (12th century), it is a valuable resource for research on Japanese culture of that time. An example of its original copy is shown in Fig. 7. Part of the *Hyohanki* has deteriorated and is missing, but all of the existing pages (comprising 2,488 diary entries) have been digitized into text format.

As described in Section 3.2.1, we need dictionaries in order to translate modern language query words into archaic words. In the case of the *Hyohanki* diary we can use some existing electronic dictionaries available on CD-ROMs. For modern Japanese we use *Kojien*, one of the most famous and comprehensive Japanese language dictionaries. For ancient Japanese we use *Kokugo-Daijiten*, which covers not only modern words but also archaic words.

Fig. 7. Part of the original copy of the historical Japanese document *Hyohanki*.

3.2.3 Experiment

We conducted a preliminary experiment to test the precision of cross-period retrieval by our proposed method. We used *Hyohanki* diary entries of as the ancient Japanese document collection and prepared three modern Japanese queries: 戦争 (war), 法要 (Buddhist service), and 裸足 (bare foot). Since the archaic equivalent of each query differs from the query itself, no relevant documents can be retrieved if we use these modern term queries. Note that we consider one diary entry as one document.

Table 2 shows the original modern Japanese query, the ancient Japanese equivalents (translations) obtained by the proposed method, and the precision of retrieval using the translations. For the queries 法要 (Buddhist service) and 裸足 (bare foot), the proposed method worked quite well: 99–100% precision (the ratio of relevant documents in retrieved documents). The query 戦争 (war), however, resulted in very poor precision (27%) because the proposed method returned two translation candidates for this query: 戦 and 軍. If we use only 戦 as the translated query we obtain 100% precision, but if we use only 軍 we obtain only 3.6% precision. This is because the archaic term 軍 has not only the meaning *war* but also meanings like *general* (officer) and *army*. The query 死亡 (death) also resulted in

Modern Japanese query	Translations	Relevant / Retrieved
戦争 (war)	軍, 戦	10 / 37 (27%)
法要 (Buddhist service)	仏事	109 / 110 (99%)
裸足 (bare foot)	跣, 裸足, 跣足	27 / 27 (100%)
死亡 (death)	没	2 / 13 (15%)

Table 2. Precision of the retrieval results in cross-period retrieval.

very poor precision (15%) because its translation 没 also means *deprivation* and *sunset*. These results suggest that we could improve the precision if we incorporate a suitable disambiguation method for the translated archaic terms. For that purpose, we could apply existing disambiguation methods used in Cross-language Information Retrieval.

4. Federated searching system for humanities databases using automatic metadata mapping

This section provides a summary of our approach to constructing a federated searching system for Japanese humanities databases using automatic metadata mapping. The goals of our system are (1) to perform metadata mapping automatically for Japanese heterogeneous humanities databases and (2) to let users access multiple humanities digital libraries by using only one query input. This section also addresses the metadata-related challenges facing Japanese humanities databases. Metadata offers library and information science a solution to the problem of describing and managing the massive quantities of explosively increasing digital information (Zeng & 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 interoperate is becoming a complex issue for research. There have been efforts to make heterogeneous standards interoperable and utilize multiple metadata standards. According to (Chan & 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. Many metadata schemas of Japanese humanities digital libraries have been accepted in terms of their semantics and content but were developed before the international metadata standards or were developed without considering the international metadata standards and 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 section therefore addresses the metadata-related challenges to constructing a federated searching system for Japanese humanities databases.

4.1 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 is 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 adopt 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 nonstandard 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.

4.2 Federated searching system for Japanese humanities databases

The conceptual architecture of our proposed federated searching system is shown in Fig. 8. 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

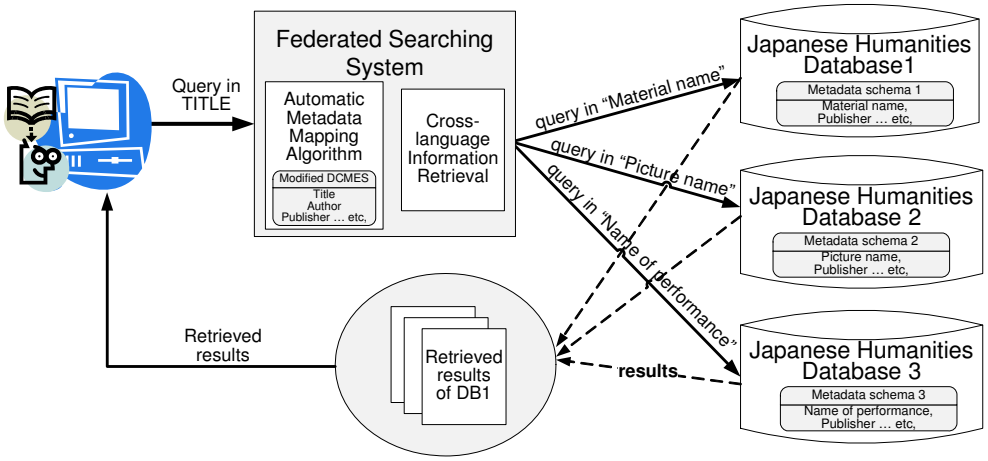


Fig. 8. Conceptual architecture of the proposed federated searching system.

interoperability or crosswalk and do not support Z39.50 protocol, Search/Retrieve Web service (SRW)/Search/Retrieve via URL (SRU), etc.
We are developing a prototype federated searching system of Japanese humanities databases—including the image database of Japanese traditional fine art Ukiyo-e, donated Japanese books database, and old Japanese books database—that are freely accessible in Japanese at the Art Research Center of Ritsumeikan University. We utilized the automatic metadata mapping method of Kimura et al. (2009). This prototype system also has a facility for cross-language searching between English and Japanese, which enables English-speaking users to search Japanese databases available only in Japanese.

4.3 Automatic metadata mapping

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 (Fig. 9) consists of two preprocessing phases and four mapping phases.

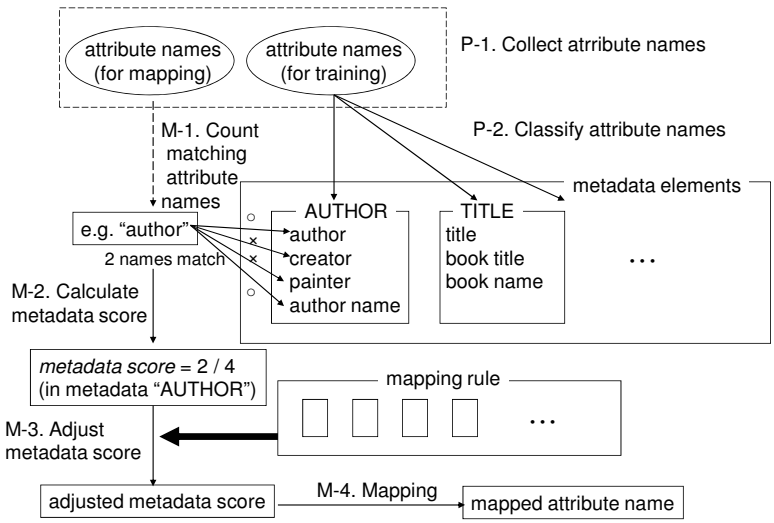


Fig. 9. Flow of automatic metadata mapping.

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.

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 *metadata 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 metadata 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 metadata score for "TEMPORAL" if the attribute name includes "year").
M-4 Map the target attribute name to the metadata element that has the highest metadata score.

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

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1, 版印2№, 版元2№, 版元2, 版元備考, 地域版, 版元統合	Character publisher, Different edition, Edition stamp #1, Publisher #1, Publisher 1, Edition stamp #2, Publisher #2, Publisher 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

Table 3. Example of results of the automatic metadata mapping.

Inspecting the data listed in Table 3, 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 element in the modified DCMES. Similarly, 11 elements were mapped to DATE, 11 to PUBLISHER, 5 to AUTHOR, and 6 to COVERAGE. These eighteen 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 depend on the combinations. Our algorithm performs automatic mapping by calculating the overall metadata scores for each metadata element, which are calculated for the words or kanji characters by using training data set and mapping rules. For instance, if the name of a metadata element has the character 名 (name), increase the metadata score for TITLE by 1, for PUBLISHER" by 0.5, and for AUTHOR by 1.

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 metadata schemas of Japanese humanities digital libraries are and that is vital to perform metadata mapping automatically.

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

Table 4. Mapping precision of the automatic metadata mapping method.

Metadata Sets	Conditions	Average precision (%)
Standard Dublin Core Metadata Element Set	Without mapping rules	73.8
Standard Dublin Core Metadata Element Set	With mapping rules	79.0
Modified Dublin Core Metadata Element Set	With mapping rules	94.9

Table 5. Comparison of metadata mapping precision.

According to the judgement of a native Japanese speaker experienced in Japanese humanities digital databases who 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 4).

The average precisions we obtained using standard DCMES without the mapping rules, using standard DCMES with the mapping rules, and using modified DCMES with the mapping rules are listed in Table 5, where one sees that the mapping precision obtained using modified DCMES with the mapping rules is 21.1 percentage points higher than that obtained using standard DCMES without the mapping rules, and this shows that 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 with mapping rules, and this shows that the modified DCMES also improves the metadata mapping considerably.

4.4 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 three collections for the sample query 風流 (elegance) in the TITLE metadata fields are shown in Fig. 10. Retrieval with other sample queries was also successful.

Fig. 10. Retrieval results obtained from three from Japanese humanities digital libraries when using automatic metadata mapping.

4.5 Retrieval in a federated searching system using English queries

Our federated searching system also retrieves resources from Japanese collections when an English query is used. This feature is very useful for users who do not understand Japanese,

and it allows searching and browsing Japanese digital libraries in English through a single interface and a single query (Batjargal et al., 2010c). We applied this feature to the Ukiyo-e image database of the Art Research Center of Ritsumeikan University, which is freely accessible in Japanese.

Ukiyo-e, Japanese traditional woodblock printing is known world-wide as one of the fine arts of the Edo period (1603–1868). The texts of Ukiyo-e databases contain archaic Japanese words which reflect the Japanese language of the Edo period. Besides providing information about Ukiyo-e prints, the Ukiyo-e database of the Art Research Center of Ritsumeikan University contains information about the content of the prints. For instance, if the subject of an Ukiyo-e print is Kabuki, the highly stylized classical Japanese dance-drama, the database contains some additional information. Sometimes explanations of cultural and social meaning for the print are also included.

67 metadata elements of the Ukiyo-e database are mapped to the modified DCMES using our automatic metadata mapping method. As shown in Fig. 11, the Ukiyo-e artist name *Kuniyoshi* as an input query was translated as 国芳 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 *Utagawa Kuniyoshi*, was treated as 歌川 (Utagawa) and 国芳 (Kuniyoshi) but not as 歌川国芳. As illustrated in Fig. 11, 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 are translated and displayed in English.

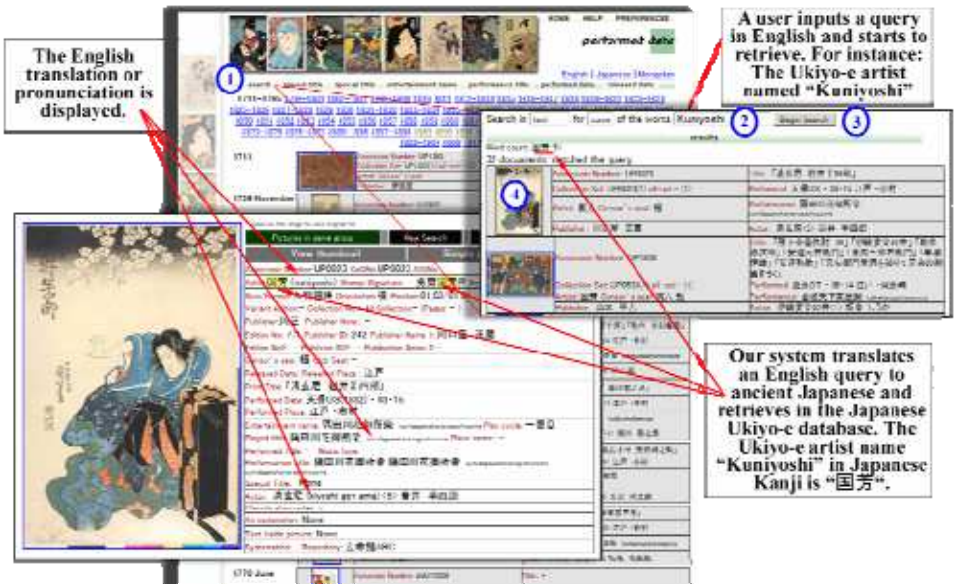


Fig. 11. Using an English query to search Japanese Ukiyo-e databases.

5. Summary

In this chapter we presented some of our work related to integrated information access technology for digital libraries. We developed technologies providing information access across different languages, periods, and cultures. These technologies will be particularly important for large digital library collections that include contents written in different languages and spanning a wide range of periods and diverse cultures. The systems presented in this chapter were developed primarily for humanities researchers but might also be useful to ordinary users because much of the knowledge and wisdom in old documents is not available in modern-language documents.

6. Acknowledgements

This work was supported in part by the Grant-in-Aid for the Global COE Program "Digital Humanities Center for Japanese Arts and Cultures (DH-JAC)" from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, MEXT Grant-in-Aid for Strategic Formation of Research Infrastructure for Private University "Sharing of Research Resources by Digitization and Utilization of Art and Cultural Materials"(Grant Number: S0991041), and MEXT Grant-in-Aid for Young Scientists (B) "Research on Information Access across Languages, Periods, and Cultures" (Leader: Akira Maeda, Grant Number: 21700271).

7. References

Batjargal, B.; Khaltarkhuu, G.; Kimura, F & Maeda, A. (2010a). An Ancient-to-modern Information Retrieval for Digital Collections of Traditional Mongolian Script. *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries (ICADL2010)*, pp. 25–28, ISBN 978-3-642-13653-5, Gold Coast, Australia, June 2010, Springer-Verlag, Berlin Heidelberg

Batjargal, B.; Khaltarkhuu, G.; Kimura, F & Maeda, A. (2010b). An Approach to Ancient-to-modern and Cross-script Information Access for Traditional Mongolian Historical Collections. *Conference Abstracts of Digital Humanities 2010*, pp. 279–282, ISBN 978-0-9565793-0-0, London, UK, July 2010, Centre for Computing in the Humanities, King's College London

Batjargal, B.; Kimura, F. & Maeda, A. (2010c). 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)*, pp. 518-521, ISBN 978-3-642-15463-8, Glasgow, UK, September 2010, Springer-Verlag, Berlin Heidelberg

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 Magazine*, vol. 12, no. 6, (June 2006), ISSN 1082-9873

Choimaa, S. & Shagdarsuren, T. (2002*). Qad-un ündüsün quriyangyui altan tobči, (Textological Study)*, Volume I, Centre for Mongol Studies, National University of Mongolia, ISBN 99929-0-119-5, Ulaanbaatar (in Mongolian)

Ernst-Gerlach, A. & Fuhr, N. (2007). Retrieval in text collections with historic spelling using linguistic and spelling variants. *Proceedings of the 7th ACM/IEEE Joint Conference on*

Digital Libraries, pp. 333–341, ISBN 978-1-59593-644-8, Vancouver, British Columbia, Canada, June 2007, ACM , New York, USA

Gotscharek, A.; Neumann, A.; Reffle, U.; Ringlstetter, C. & Schulz, K. (2009). Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. *Proceedings of the 3rd Workshop on Analytics for Noisy Unstructured Text Data*, pp. 69–76, ISBN 978-1-60558-496-6, Barcelona, Spain, July 2009, ACM, New York, USA

Hauser, A.; Heller, M.; Leiss, E.; Schulz, K. & Wanzeck, C. (2007) Information Access to Historical Documents from the Early New High German Period. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, January 2007

Kawashima, M.; Akama, R.; Yano, K.; Hachimura, K. & Inaba, M. (2009). *New Directions in Digital Humanities for Japanese Arts and Cultures*. Nakanishiya Shuppan, ISBN 978-4-7795-0324-5, Kyoto

Khaltarkhuu, G. & Maeda, A. (2006). Retrieval Technique with the Modern Mongolian Query on Traditional Mongolian Text. *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL2006)*, pp. 478–481, ISBN: 3-540-49375-1, Kyoto, Japan, November 2006, Springer-Verlag, Berlin Heidelberg

Khaltarkhuu, G. & Maeda, A. (2007). Building a Digital Library of Traditional Mongolian Historical Documents. *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries (JCDL2007)*, p. 483, ISBN 978-1-59593-644-8, Vancouver, Canada, June 2007, Springer-Verlag, Berlin Heidelberg

Khaltarkhuu, G. & Maeda, A. (2008). Developing a Traditional Mongolian Script Digital Library. *Proceedings of the 11th International Conference on Asia-Pacific Digital Libraries: Universal and Ubiquitous Access to Information*, LNCS, vol. 5362, pp. 41–50, ISBN 978-3-540-89532-9, Bali, Indonesia, December 2008, Springer-Verlag, Berlin Heidelberg

Kimura, F. & Maeda, A. (2009). An Approach to Information Access and Knowledge Discovery from Historical Documents. *Conference Abstracts of the Digital Humanities 2009(DH09)*, pp. 359–361, ISBN 978-061-52-9929-7, College Park, MD, June 2009, Maryland Institute for Technology in the Humanities

Kimura, F.; Toba, T.; Tezuka, T. & Maeda, A. (2008). Federated Searching System for Humanities Databases Using Automatic Metadata Mapping. *Proceedings of 9th International Conference on Dublin Core and Metadata Applications*, pp. 139–140, ISSN 1939-1366, Seoul, Korea, October 2009, the Dublin Core Metadata Initiative

Kitamura, M. & Matsumoto, Y. (1996). Automatic Extraction of Word Sequence Correspondences in Parallel Corpora, Proceedings of the 4th Workshop on Very Large Corpora, pp. 79–87

Koolen, M.; Adriaans, F.; Kamps, J. & Rijke, M. (2006). A Cross-Language Approach to Historic Document Retrieval. *Proceedings of the 28th European Conference on IR Research: Advances in Information Retrieval*, LNCS, vol. 3936, pp. 407–419, ISBN 978-3-540-33347-0, London, UK, April 2006, Springer-Verlag, Berlin Heidelberg

Pilz, T.; Ernst-Gerlach, A.; Kempken, S.; Rayson, P. & Archer, D. (2008). The Identification of Spelling Variants in English and German Historical Texts: Manual or Automatic. *Literary and Linguistic Computing*, vol. 23, no. 1, November 2008, pp. 65–72, ALLC, ACH, Oxford University Press, ISSN 0268-1145, Oxford, UK

Shagdarsuren, T. (2001). *Study of Mongolian Scripts (Graphic Study of Grammatology)*, Enlarged second edition. ISBN 99929-5-347-0, Urlakh Erdem Khevleliin Gazar, Ulan Bator (in Mongolian)

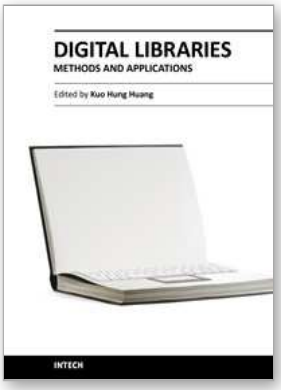
Tanaka, T. (2002). Measuring the Similarity between Compound Nouns in Different Languages Using Non-Parallel Corpora, Proceedings of the 19th COLING, pp. 981–987, ISBN 978-1-55860-896-2, Morgan Kaufmann Publishers, December 2002, Taipei, Taiwan

The Dublin Core Metadata Initiative (2008). Dublin Core Metadata Element Set, Version 1.1. http://dublincore.org/documents/dces/

Tsevel, Y. (1966). *Mongol helnii tovch tailbar toli*. Ulsiin Khevleliin khereg erkhlekh Khoroo, Ulan Bator (in Mongolian)

Tungalag, D. (2005). *Mongol ulsiin undesnii nomiin san dahi Mongoliin tuuhiin gar bichmeliin nomzuin sudalgaa*, Volume 1. ISBN 99929-6-313-1, Time Printing, Ulan Bator (in Mongolian)

Zeng, M. L. & Jian, Q. (2008). *Metadata*, ISBN 978-1-85604-655-8, Facet Publishing, London



Digital Libraries - Methods and Applications

Edited by Dr. Kuo Hung Huang

ISBN 978-953-307-203-6
Hard cover, 220 pages
Publisher InTech
Published online 04, April, 2011
Published in print edition April, 2011

Digital library is commonly seen as a type of information retrieval system which stores and accesses digital content remotely via computer networks. However, the vision of digital libraries is not limited to technology or management, but user experience. This book is an attempt to share the practical experiences of solutions to the operation of digital libraries. To indicate interdisciplinary routes towards successful applications, the chapters in this book explore the implication of digital libraries from the perspectives of design, operation, and promotion. Without common agreement on a broadly accepted model of digital libraries, authors from diverse fields seek to develop theories and empirical investigations that to advance our understanding of digital libraries.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821