International Journal of Advances in Computer Science & Its Applications
Volume 6 : Issue 1 [ISSN 2250-3765]
Publication Date : 18 April, 2016

Design and Prototype Implementation of a Federated Search System for Multiple Japanese Humanities Databases

Biligsaikhan Batjargal

***Abstract*— This paper discusses the authors' approach in implementing a Federated Search System for multiple Japanese humanities databases. We introduce a prototype federated search system for online Japanese databases, which retrieves multiple and heterogeneous humanities databases in parallel and provides real-time integration of the results. The authors' results are explored further in the paper.**

***Keywords*— Multilingual digital archive, federated search, Japanese humanities databases, information retrieval.**

I. Introduction

As a result of worldwide digitization over the last decade, many humanities and cultural institutions started to expose digital objects on the Internet in a variety of languages through a diversity of interfaces. Likewise, the Art Research Center (ARC) at Ritsumeikan University has been constructing digital archives of tangible or intangible heritages of Japanese arts and cultures. Many traditional or contemporary cultural heritages of Japan are being digitized and made publicly available as searchable databases. The databases at the ARC are designed, created, maintained and updated by individual humanities scholars and researchers of different fields. Thus, it has become apparent that the databases at the ARC are heterogeneous and different from each other.

With increasingly more heterogeneous databases becoming available on the Internet, it is essential to develop methods for accessing these databases of vast and valuable collections of cultural heritage easily and thoroughly. Users need an efficient way of searching multiple databases. Therefore, we aimed to develop an easy-to-use and extensible system for users to allow searching and browsing multiple heterogeneous Japanese humanities databases through a single interface and a single query. Here, we present our approach to solve the challenges for realizing such a system, which accesses to diverse data across multiple Japanese humanities databases. We applied a federated search technology to the public version of 1) ARC Ukiyo-e Portal Database (approximately 28,000 prints), 2) ARC Early Japanese Book Database (approximately 16,000 books), and 3) ARC Japanese modern book database (approximately 400 books) among other 40 different databases of the ARC, which is freely accessible in Japanese.

Biligsaikhan Batjargal
Research Organization of Science and Engineering / Ritsumeikan University
Japan

II. Related Work

The prototype system that have introduced in this paper belongs to the *federated search* (a.k.a *distributed information retrieval*) [1] paradigm. There are other approaches for accessing multiple databases i.e., *harvesting* or *web crawling*, which collect data from various sources in advance. Usually, such approaches use a master index that requires frequent updates and massive storage. Without frequent updates, information contained in each database becomes outdated so that new updates would be unavailable for searching. However, this paper concerns the real-time access of up-to-date contents of multiple databases.

There are several federated search systems such as OpenSiteSearch [2], JAFER [3], Sesat [4], JzKit3 [5], Xerxes [6] and Pazpar2 [7]. By employing the Pazpar2, we have developed a federated search system for accessing and retrieving multiple databases in parallel, and providing real-time integration of the search results. The following section discusses the current implementation of the proposed prototype system.

III. The Proposed System

The main features of the proposed system are: 1) to let users access multiple Japanese humanities databases in parallel through a single interface and a single query, 2) to aggregate, integrate and display the retrieved results in a single interface instantaneously. The overall design of the proposed system is illustrated in Figure 1. We believe any traditional practice of accessing each individual database separately could be improved by using the proposed system. Using the proposed system, users can get better results than the native database's search interface. Even though, the same content is being searched, the proposed system enhances the retrieved results, so that users are provided with faceted navigations and unified search results of bilingual data holding basic information, and links where the original records along with detailed information can be found.

A. Simultaneous access to multiple humanities databases at the ARC

Federated search technology enables users to search multiple databases simultaneously through a single query input [1]. Users can view search results in a single integrated list. In other words, users do not need to search each databases individually.

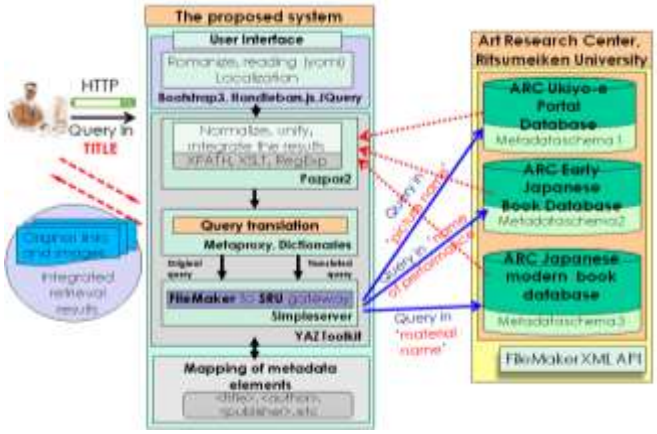

Figure 1. Conceptual architecture of the proposed system.

We adopted Search/Retrieve via URL (SRU) based approach for searching multiple databases by sending queries in real time. SRU is a set of standard synchronous search protocols and standards for Internet search. The Library of Congress of the United States serves as the maintenance agency for these standards and protocols [8].

As illustrated in Figure 1, the proposed federated search system is able to retrieve multiple and heterogeneous back-end Japanese databases [9]. Certain databases at the ARC are made available via a gateway using the federated search protocol SRU and can be searched and retrieved simultaneously. An SRU gateway accesses to the databases in the FileMaker format via Custom Web Publishing with XML API [10] and return a result set represented in SRU-ready XML, when a search request is received from the users. We have utilized a Perl module Net::Z3950::SimpleServer [11] along with YAZ toolkit [12] for listening to incoming connections, and implementing the SRU protocol.

In order to convert Common Query Language (CQL) [13] query, which is used by SRU, into FileMaker query [10], we have utilized a Perl module CQL::Parser [14]. We have employed a metasearching middleware Pazpar2 [15] for searching multiple databases in parallel and providing on-the-fly normalization and integration of the search results from diverse and multiple databases.

We have also employed our previous approach [16] for mapping the diverse metadata schemas. By using this approach, if a user wants to find a resource with the query word in the title, the system retrieves resources having the query word in the title or any metadata field that is similar to the title or could be treated as a title (e.g., “picture names”, “material names”, “book title”, etc.) and retrieves those resources from diverse databases. Although, the records returned from multiple databases can be in different metadata schemas, a simple schema is chosen at the front-end for integrating returned results while utilizing other metadata elements without omission. A simple metadata schema similar to Dublin Core Metadata Element Set [17] is chosen and basic elements such as <title>, <author>, <subject>, <date>, <description>, <publisher>, <type>, <identifier>, <source>, <coverage> and <rights> are adopted for displaying returned records as an integrated result listing in a single interface.

B. Data returned in the native interfaces of databases

Sample search results of 1) ARC Early Japanese Book Database, 2) ARC Japanese modern book database, and 3) ARC Ukiyo-e Portal Database are shown in Figure 2, Figure 3 and Figure 4, respectively. The above databases can be searched and retrieved simultaneously via the gateway that we have explained in Section III.A.

C. Bilingual queries over Japanese databases

Additionally, a feature of cross-language searching between English and Japanese [18] have integrated to the prototype system for enabling English-speaking users to search databases available in Japanese by using English queries. Such a feature is very useful, since the databases in Japanese institutions are mostly available in Japanese, so that users who do not understand Japanese may not find the desired information.

A dictionary-based query translation approach is adopted by utilizing a domain-specific dictionary, which contains the terms related to Japanese arts and cultures. The proposed feature works well with a variety of keywords (i.e. no full sentences) that may include the personal names, specific terms such as “Geisha”- traditional Japanese female entertainers, “Fuji” and “Sumo”- Japanese traditional wrestling. For instance, if the search query submitted by the user is a name of the Ukiyo-e artist i.e., “Utagawa Hiroshige”, then the query “Utagawa” will be translated into Japanese as “歌川広重” and sent to multiple Japanese databases. Such a cross-language search feature will contribute to expand the understanding of Japanese arts and cultures in the English-speaking world. Moreover, it would help users to involve a further research in Japanese arts and cultures that the English-speaking user may or may not understand.

In this prototype, the users’ query and its translation will be sent to the same database simultaneously and the retrieved results will be merged by utilizing a metasearching tool Metaproxy. Metaproxy is a proxy front end server that presents a single SRU front end to multiple back end database servers. Metaproxy combines caching, load balancing, filtering, and cross database merging of the result set. It is designed for integrating multiple back end databases - within a single organization or across multiple organizations - into what appears from the user’s point of view to be a single searchable resource [19].

D. Normalizing and generating bilingual data

After gathering data from multiple databases, the proposed system performs two additional steps, which are 1) generating

bilingual data; and 2) normalizing data, merging and displaying results in a user-friendly way.

Figure 2. Search results of ARC Early Japanese Book Database.

Figure 3. Search results of ARC Japanese modern book database.

Figure 4. Search results of ARC Ukiyo-e Portal Database.

1) Generating bilingual data

This section discusses how to make bilingual data from the returned search results of multiple humanities databases.

Knowing the exact pronunciation or *yomi*, which gives phonetic representation of the certain word, is helpful for users who do not understand Japanese language. Therefore, attaching the *yomi* or transcriptions in web pages is useful because of the kanji's different readings for a given context. Usually, *yomi* is written in *kana*, although *romaji* –romanized representation of Japanese is also used in the West. Some of the databases use transcriptions in *romaji*. Therefore, in the proposed system the following features also have been included: 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 users may understand Japanese kanji or *kana*, but *romaji* might be more readable for others. *Yomi* might be needed to know the correct pronunciation.

At first, the proposed system searches the *yomi* and translations from the returned search results. If the *yomi* or transcription is not found, the proposed system calls MeCab – Part-of-speech and morphological analyzer for Japanese to generate *yomi* for the Japanese text. A custom dictionary with the terms, special terminology and names of the artists, etc., which are related to Japanese arts and culture, was added to the MeCab and used in conjunction with the MeCab's default dictionaries. The default dictionaries are not sufficient to provide all the *yomi* of Japanese arts and culture related terms.

After producing the bilingual data using the *yomi*, transcriptions and translations, our system will parse, normalize and aggregate the search results and display them in a user-friendly way. Users are provided with a list of bilingual data, holding basic information and a link of the original

records along with detailed information. As shown in Figure.5, transcriptions in romaji, English title, and the translation of the Japanese title are used for displaying Japanese content in English pages. Meanwhile, as shown in Figure 6, transcriptions in kana, Japanese title will be displayed in Japanese pages. A button is provided to toggle between English and Japanese text by letting users quickly change the language of the search results.

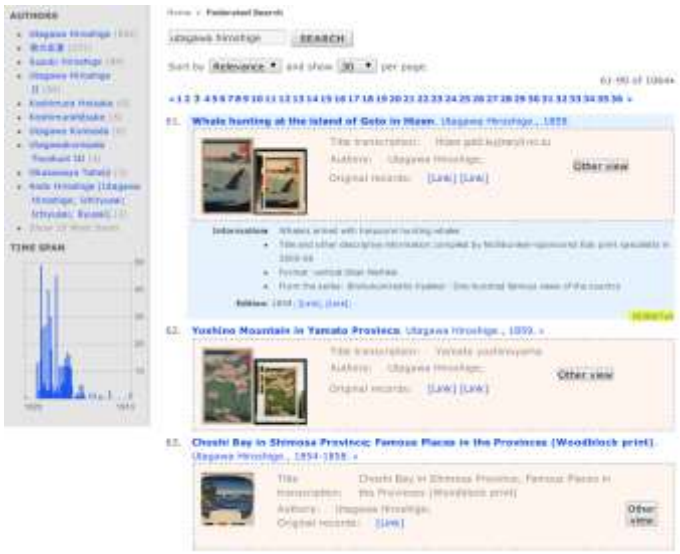

Figure 5. Search results with English data.

Figure 6. Search results with Japanese data.

Furthermore, faceted interfaces offer a new way of finding and browsing databases as an addition to the typical keyword searching and browsing. In the proposed system, a facet generation task has been implemented and applied to authors, genre and subject fields.

Unifying into a common representation allows users to browse diverse contents of various databases in a single page. Such a feature is helpful for users to find the necessary records easily by avoiding clicking each different representation. In general, using the proposed approach, users can get better results. Even though, the same content is being searched, the proposed system improves the facets, so that users are provided with a better result, where the diverse representations are unified and irrelevant data are eliminated.

2) Normalizing

In this section, normalizing the data for sorting, merging and displaying the unified results is discussed.

First of all, all full-width (zenkaku) alphanumeric characters should be converted to ASCII in order to normalize data and perform further tasks such as merging, relevance ranking, sorting, and showing faceted results. Data returned by multiple databases are diverse due to result presentation styles. Metadata values may have an inconsistent format between databases. For instance, some unnormalized or non-English data such as “about 1854 − 60 (Ansei era)”, “Utagawa Hiroshige (歌川広重) (Print artist)” and “歌川広重(初代)” might be obtained. Author names could be represented in a variety of ways within each database. As shown in Table 1, the name of the Ukiyo-e artist “Utagawa Hiroshige” could be represented differently. Therefore, some tasks for unifying such diverse representations are necessary. Some databases have author names written as “First name, Last name” format, which faceting or merging algorithms will treat differently than ‘Last name’ ‘Fist name’. Moreover, as can be seen in Table 1, some of the representations have included birth-and-death dates, after the artist name. In order to show better results, birth and death dates of artists need to be removed and the names converted into the more common form i.e. ‘Last name’ ‘Fist name’ while capitalizing the first letter of the first name and last name or converting other letters to lower-case.

TABLE I. UTAGAWA HIROSHIGE REPRESENTED IN VARIOUS WAYS

No	Naming	Rules that can be applied
1	Ando Hiroshige [Utagawa Hiroshige; Ichiryusai; Ichiyusai; Ryusai] (Japanese, 1797–1858)	-splitting <author>, -removing birth and death dates
2	Andô Hiroshige (Japanese, 1797–1858)	-removing birth and death dates
3	Utagawa Hiroshige (Japanese, 1797–1858)	
4	Hiroshige	none
5	Utagawa Hiroshige	
6	Ando Hiroshige	
7	Utagawa Hiroshige I	
8	Hiroshige, Utagawa	-reversing the "Last name, First name"
9	Utagawa Hiroshige Japanese, 1797–1858	-removing birth and death dates
10	Utagawa Hiroshige (歌川広重)	-language separation
11	歌川広重	-romanization
12	歌川広重(初代)	-romanization, -converting the generation
13	広重〈1〉	

Another consideration is, a date may be represented as “1857” in a record, and as “about 1854−60 (Ansei era)” in another and further as “安政 4 年” in Japanese. The format of

a date field has to be normalized across all results from all sources. Japanese calendar years have to be converted into Gregorian calendar years by using a simple hash map. Regular expressions could be better candidates for performing the above tasks efficiently. A sample of regular expressions is shown in Table 2.

TABLE II. SAMPLE OF REGULAR EXPRESSIONS FOR NORMALIZING ARTIST NAMES

No	Rules	Example	Regular expression
1	splitting <author>	Utagawa Hiroshige; Ando Hiroshige	;\s*
2	removing birth and death dates	Utagawa Hiroshige (Japanese, 1797–1858), Utagawa Hiroshige Japanese, 1797–1858	\(?Japanese?, \s*[0-9-]+$
3	reversing the last name, first name	Hiroshige, Utagawa	s/(.*),\s*(.*)/2 $1
4	choosing a date from a string	about 1854–60 (Ansei era)	[^0-9\.]+ or .*?(([0-9]*\.) ?[0-9]+).* $1

E. User Inteface

As the front-end web user interface, we use the combination of Bootstrap 3 [20], Handlebars.js [21] and jQuery [22] along with asynchronous JavaScript and XML (AJAX) techniques. The user interface communicates with SRU servers via Pazpar2 to display the results dynamically and allows the user to interact with the retrieved results that exchanged asynchronously between browser and SRU servers. Open-source engines Handlebars.js, jQuery and AJAX techniques change the content dynamically without reloading the entire search results from multiple servers. In short, the front-end web user interface sends users' query to and retrieves search results from SRU servers asynchronously by using the above techniques.

IV. Summary and discussions

In this paper, a prototype system for retrieving multiple Japanese databases in parallel and integrating retrieved results instantaneously has been introduced. Such a system along with a cross-language searching feature is useful for users, who do not understand Japanese, and it allows searching and browsing Japanese multiple humanities databases in a single interface so that it helps users to save time. Cross-language searching feature is useful for humanities researchers who are looking for relevant materials written in Japanese. For instance, for someone conducting research in Japanese arts and culture, accessing databases described in Japanese is often needed to understand the elements of Japanese arts and cultures.

Several challenges in dealing with heterogeneous data across multiple databases, such as data extraction, generation and normalization of bilingual data have been discussed.

As a future work, a humanities-researchers-centered evaluation will be conducted by experts and humanities researchers on usefulness of the proposed system by doing usability surveys. Further improvements will be done based on the evaluation results and user feedback. We are also planning to apply the proposed system to other humanities databases.

References

[1] M. Shokouhi, and S. Luo, "Federated Search," Foundations and Trends in Information Retrieval, vol. 5 (1), pp. 1–102, 2011.

[2] Bjorgensen and others, "OpenSiteSearch," Accessed: December 6, 2015. http://opensitesearch.sourceforge.net/.

[3] J. D. Matthew, C. Tatham, and A. Corfield, "JAFER Toolkit Project," in VINE 35, no.1/2, 2005, pp. 49–51. doi: 10.1108/03055720510588461.

[4] Schibsted ASA, "Sesat - SEsam Search Application Toolkit," Accessed: December 6, 2015. http://sesat.no/.

[5] Knowledge Integration Ltd, "JZKit 3," Accessed: December 6, 2015. http://www.k-int.com/jzkit.

[6] D. Walker, "xerxes-potal," Accessed: December 6, 2015. http://code.google.com/p/xerxes-portal/.

[7] IndexData, "Pazpar2," Accessed: December 6, 2015. http://www.indexdata.com/pazpar2.

[8] The Library of Congress, "Search/Retrieve via URL," Accessed: December 6, 2015. http://www.loc.gov/standards/sru/.

[9] B. Batjargal, F. Kimura, and A. Maeda, "Metadata-related Challenges for Realizing a Federated Searching System for Japanese Humanities Databases," In Proc. of the 11th Int. Conf. on Dublin Core and Metadata Applications, The Hague, pp. 80–85, September 2011.

[10] FileMaker Inc, "FileMaker Server 14. Custom Web Publishing Guide," Accessed: December 6, 2015, https://fmhelp.filemaker.com/docs/14/en/fms14_cwp_guide.pdf.

[11] M. Taylor, "Simple Perl API for building Z39.50 servers," Accessed: December 6, 2015, http://search.cpan.org/~mirk/Net-Z3950-SimpleServer/SimpleServer.pm.

[12] IndexData, "YAZ toolkit," Accessed: December 6, 2015, http://www.indexdata.com/yaz.

[13] The Library of Congress, "The Contextual Query Language," Accessed: December 6, 2015, http://www.loc.gov/standards/sru/cql/.

[14] E. Summers, B. Cassidy, and W. Hengst, "CQL::Parser," Accessed: December 6, 2015, http://search.cpan.org/~bricas/CQL-Parser-1.13/lib/CQL/Parser.pm.

[15] IndexData, "Pazpar2," Accessed: December 6, 2015, http://indexdata.com/pazpar2/.

[16] The Dublin Core Metadata Initiative, "Dublin Core Metadata Element Set, Version 1.1," Accessed: December 6, 2015. http://dublincore.org/documents/dces/.

[17] B. Batjargal, F. Kimura, and A. Maeda, "Approach to cross-language retrieval for Japanese traditional fine art: Ukiyo-e database," In Proc. of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2010), UK, 2010, pp. 518–521, Springer Berlin Heidelberg, doi: 10.1007/978-3-642-15464-5_71.

[18] B. Batjargal, F. Kimura, and A. Maeda, "Providing universal access to Japanese humanities digital libraries: an approach to federated searching system using automatic metadata mapping," Journal of Zhejiang University SCIENCE C vol. 11, no. 11, pp. 837–843. November 2010.

[19] IndexData, "Metaproxy," Accessed: December 6, 2015, http://www.indexdata.com/metaproxy.

[20] "Bootstrap 3," Accessed: December 6, 2015, http://getbootstrap.com/.

[21] "Handlebars.js," Accessed: December 6, 2015, http://handlebarsjs.com/.

[22] The jQuery Foundation, "jQuery," Accessed: December 6, 2015, https://jquery.com/.

About Author:

Doctor of Engineering, Ritsumeikan University, Japan.
Senior researcher at the Research Organization of Science and Engineering, Ritsumeikan University, Japan.
Interested in digital libraries, digital humanities and multilingual information retrieval.