Multilingual Access to Dublin Core Metadata of ULIS Library

Danyang Wen, Tetsuo Sakaguchi, Shigeo Sugimoto, and Koichi Tabata
University of Library and Information Science
1-2 Kasuga, Tsukuba, Ibaraki 305-8550, Japan
{wdy, saka, sugimoto, tabata}@ulis.ac.jp

Abstract

With the recent Internet expansion, persons all over the world can access more and more document databases. As Unicode has become more popular, the environment for multilingual retrieval has been improved to some extent. However, there are still numerous problems to be solved, such as the multilingual input and display. This paper proposes a system for retrieving ULIS Japanese metadata. The system provides convenience for the overseas users for multilingual access by solving these problems and by indicating candidates for translated words and Boolean queries based on the statistics of the system's behavior.

Keywords: *Multilingual Access, Dublin Core Metadata, Cross-Language Information Retrieval*

1. Introduction

Widespread Internet improvements have resulted in voluminous data being stored in libraries and museums and being accessed by overseas users. The needs for multilingual access are increasing more and more in the actual world. Cross-Language Information Retrieval (CLIR)[1][2] is the most representative research in the multilingual retrieval field. CLIR began in the 1960's and was first used for map and road indexing systems. In the 1970's, there was the Salton experiment [3]. The most basic concept in this field is that users can find the useful data from foreign language database using their mother tongue.

Overseas users wanting to use a Japanese database usually face many problems. The first problem is often how to display and input Japanese words. With the spread of Unicode, this problem has been solved to some extent. However, in actual cases, overseas users often cannot read Japanese on their display because the Japanese character font is not installed. Often the overseas user doesn't know how to input Japanese from the keyboard. Furthermore, the users generally don't know Japanese well and feel it is difficult to find the proper keyword, which is very important for retrieval work. Most users may not finish their retrieval unless they are working in their mother language.

ULIS metadata are metadata records that are collected and created by the University of Library and Information Science (ULIS). The major resources are WWW documents contained in the WWW sites of libraries and related institutions worldwide. There are now about 30 thousand records in the ULIS metadata.

The subject gateway service is the primary function of ULIS-DL(digital library). We collect resources and create metadata records of the resources. The major resources are WWW documents contained in the WWW site of library and LIS (Library Information Science)-related institutions. The metadata elements set, called ULIS Core (UC), is defined based on the 15 Simple Dublin Core elements with small extensions. The extensions are character code and country of publication as primary elements, and pronunciation information as a sub-element of every element. The metadata records are basically created for every document object collected by a

crawler.

This paper proposes a method of Multilingual access to ULIS Japanese metadata so people all over the world can access the data. This system is mainly intended for those who can understand Japanese and have problem to read Japanese fonts or input Japanese keywords by computer. For example although there are some same characters in Japanese and Chinese, but they have the different rule to input the character for computer inputting. Chinese people can't read or input any Japanese word without installing the fonts and inputting software.

This system uses Multilingual HTML (MHTML) technology [4][5] to solve the problem of displaying multilingual characters. MHTML is a technology developed by the authors themselves several yeas ago. Using the technology, multilingual documents can be browsed on an off-the-shelf Web browser, and a multilingual gateway service to browse foreign documents is provided.

To solve the problem of Japanese input in this system, an overseas user chooses a Japanese word from a list of candidate words shown on his display by clicking with a mouse.

In addition, the following two methods are introduced to support the user who is not so familiar with Japanese.

(1) Support to find proper Japanese words.

In order to get a good retrieval result, users have to input the proper keyword. Overseas users may have difficulty finding the proper Japanese keyword for their retrieval. Based on statistics of user's behavior in the system, candidates of translated Japanese words are shown to new users in the translation phase.

(2) Support to find a proper Boolean query.

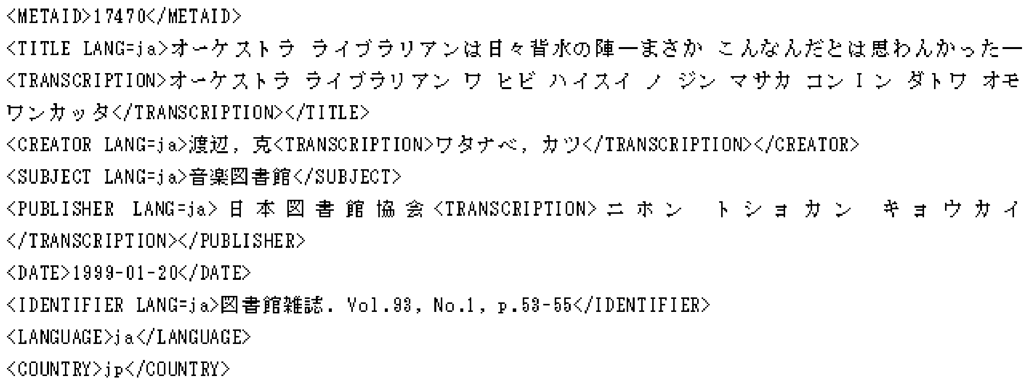
In retrieval activities, users often use "or," "and," and "not" operators to get the proper amount of retrieval results. However, it is a difficult for the users, especially foreign users, to decide the proper Boolean query. Since the foreign users often have problems understanding Japanese, they also have difficulties deciding the proper search formula. In the proposed system, users can select a search formula by referring to candidates of Boolean queries based on the statistics of the system's behavior.

In order to make the system useful for more languages, English is used as the mediator language. In the Internet world today, we find that English is, in fact, the common language. In the multilingual retrieval world, many countries begin their research from English as the first step. Also, in the digital dictionary world, there are numerous free dictionaries from English to any other language. In contrast, we have not been able to find any free any-to-any digital dictionaries. In the proposed system, a keyword in a certain language is translated into an English keyword. The English keyword is then translated into a Japanese keyword. (any foreign language --> English ---> Japanese).

2. ULIS Metadata

The principal purpose of the ULIS Digital Library is to be a subject gateway for Internet documents in the field of library and information science[6]. Its metadata records are basically created for every document automatically collected by a crawler program. Human catalogers create metadata records from the collected documents. Created metadata records are then approved by a manager.

The ULIS metadata element is defined based on the 15 Simple Dublin Core elements [7] with small extensions. The extensions are character code and country of publication as primary elements, and pronunciation information as a sub element of every element (See Fig. 1)



```
<METAID>17470</METAID>
<TITLE LANG=ja>オーケストラ ライブラリアンは日々背水の陣ーまさか こんなんだとは思わんかったー
<TRANSCRIPTION>オーケストラ ライブラリアン ワ ヒビ ハイスイ ノ ジン マサカ コンＩン ダトワ オモ
ワンカッタ</TRANSCRIPTION></TITLE>
<CREATOR LANG=ja>渡辺，克<TRANSCRIPTION>ワタナベ，カツ</TRANSCRIPTION></CREATOR>
<SUBJECT LANG=ja>音楽図書館</SUBJECT>
<PUBLISHER LANG=ja> 日 本 図 書 館 協 会 <TRANSCRIPTION> ニ ホ ン  ト シ ョ カ ン  キ ョ ウ カ イ
</TRANSCRIPTION></PUBLISHER>
<DATE>1999-01-20</DATE>
<IDENTIFIER LANG=ja>図書館雑誌．Vol.93，No.1，p.53-55</IDENTIFIER>
<LANGUAGE>ja</LANGUAGE>
<COUNTRY>jp</COUNTRY>
```

Figure 1 Example of the ULIS Metadata

3. System Outline

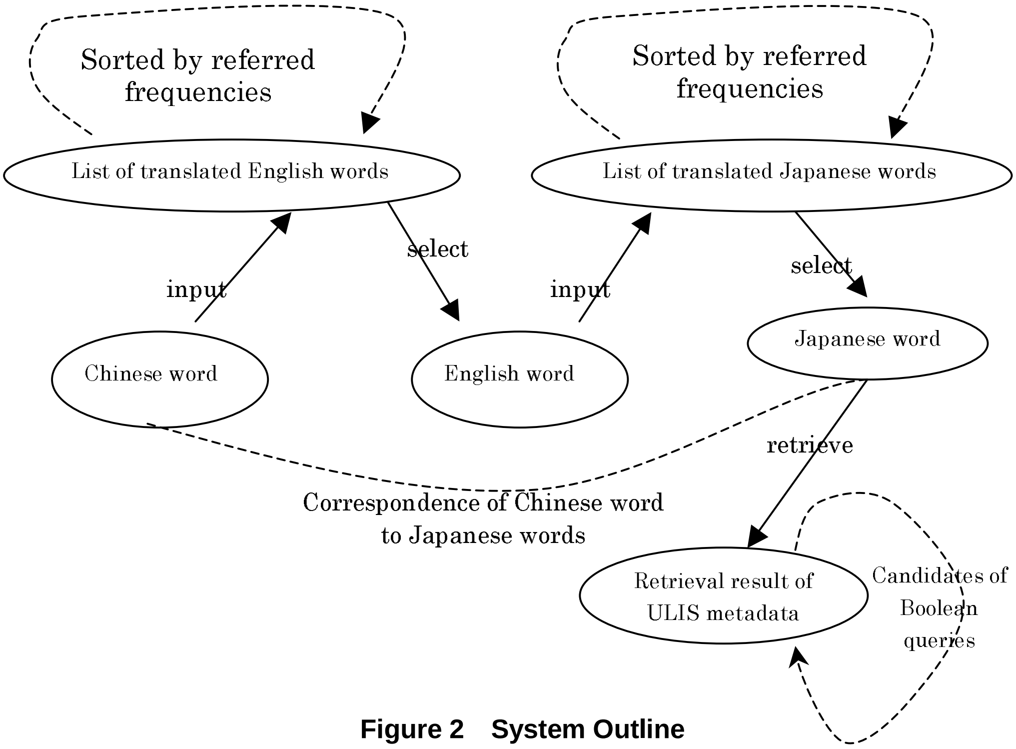


Figure 2 System Outline

In the case of Chinese, a user inputs his Chinese keyword into the system. The system first shows the list of translated English words (See Fig. 2). The user then selects a keyword from the list and inputs it to the system. The system will return a list of translated Japanese words to the user. The user next selects his keyword from the list and sends a Boolean query including it to the system. The order of listed items is always re-sorted by the referred frequencies. Finally, the system returns the retrieval result to the user.

The following functions are provided for users in the system.

(1) Selection of the translated word.

On receiving a user request, the system will search the digital dictionary, and return a list of all translated words to the user. The system can also perform a Left-Hand Truncation search as well. A user can also input two words together; the system will then return the two corresponding

results together.

The user selects a Japanese word by indicating the item number attached to it in a list of translated Japanese words.

When a user inputs his mother language or English, candidate functions supply users the past consulting information, and make sure users can easily get his needed keyword. Here is an example to smooth the search progress:

(a) When an English word is input, the system will return all of the translated Japanese words sorted by the frequencies that the words have been referred to for retrieval (Figure 3 -a). The user selects one of them by indicating its associated item number.

(b) When a Chinese word is input, the system will return all of the translated English words sorted by the frequencies that they have been referred to in the system. The user selects one of them and brings it into process (a).

At the same time, the system will also display Japanese words that consequently corresponded to the Chinese word through the process (a) and (b) in the past retrieval (Figure 3 -b). (c) When two Chinese words are input, the system will return two lists of translated English words and corresponding Japanese words (Figure 3 -c).

(2) Retrieval of the Metadata

To satisfy his or her demand, a user takes

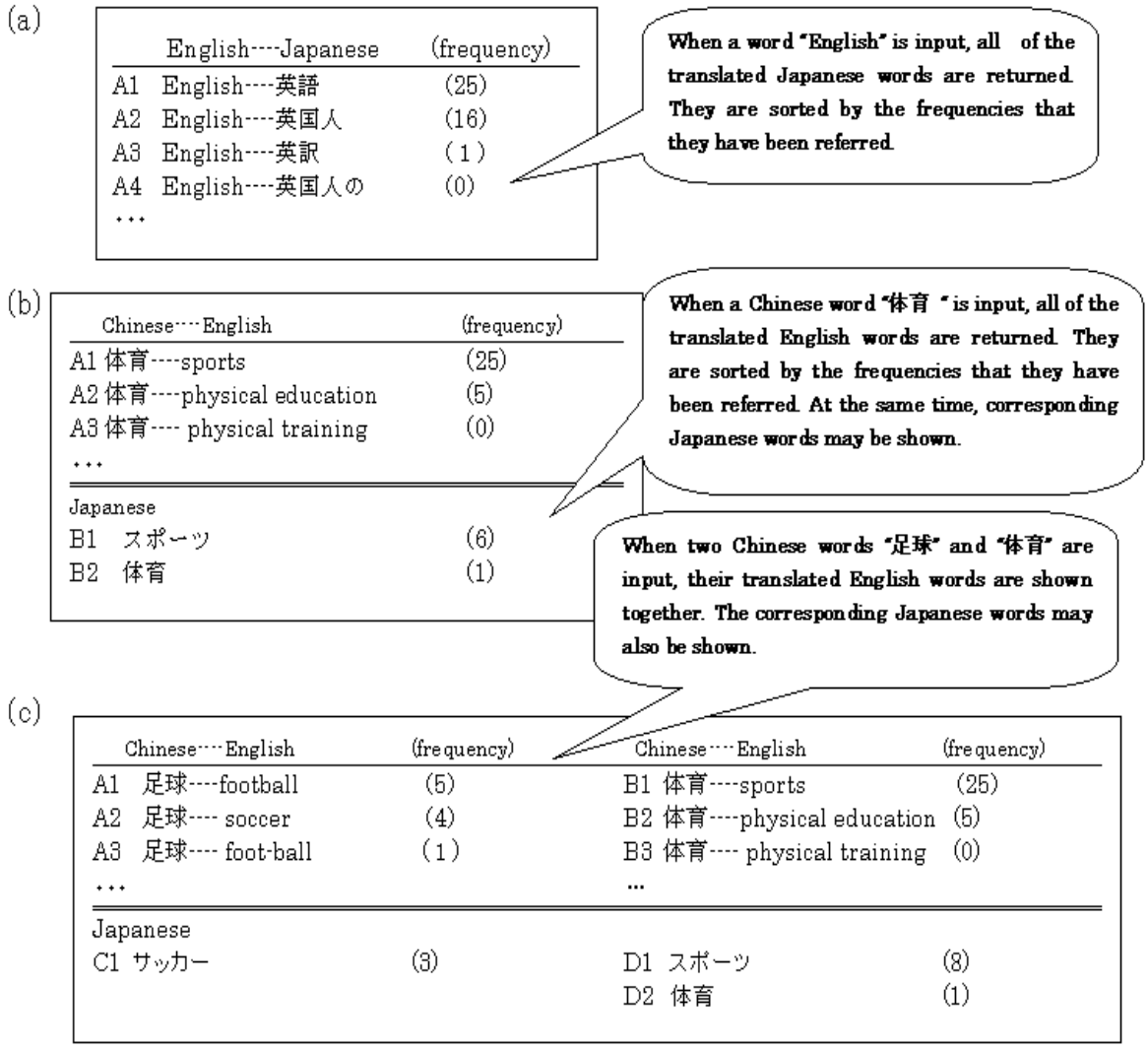


Figure 3 Lists of Candidate Translated Words

advantage of the "or" search, "and" search, and "not" search to get the proper amount of data. After considering the data result, the user may select another Boolean query for a new retrieval. The results can be shown in abbreviated style or detailed style.

In normal retrieval systems, data is often retrieved in the following steps.

(a) Input the keyword.
(b) Read the retrieval result.
(c) Change the keyword and begin a new retrieval.

In taking these steps, users may face the following problems. (a) Users may not find the proper keyword. (b) It may take a long time to select the proper result from a large number of results. (c) It may be difficult for the users to find a new keyword or select a new Boolean query. For multilingual access in particular, these problems become more serious for the users. In the proposed system, candidate functions are introduced to help the user to solve these problems.

Candidate functions play the following roles.

(i) Support inputting keywords and Boolean queries.
(ii) Facilitate understanding the quantity of the retrieval result.
(iii) Provide information on ranking the retrieval results.

When a user inputs a Boolean query, the system not only returns its result but also shows related queries performed in the past and the amounts of their results. Related queries mean all past queries including at least one keyword appearing in the current query.

4. System Implementation

This system is available on the WWW. The interface of the system is based upon the Common Gateway Interface (CGI) program and HTML form. The dictionary function, metadata retrieval function and candidate function are written in C language. The OpenText [8] search engine is utilized for searching dictionaries [9] and ULIS metadata. The candidate function is realized on a miniSQL [10] and PostgreSQL [11] DBMS server. The system uses an MHTML applet to display multilingual files (See Fig. 4). We are using EDR Electronic Dictionary for translating English to Japanese, and using free Chinese English dictionary with 35,000 words which was made by Mathias Johansson to translate Chinese to English.

(1) OpenText search engine

OpenText is a full-text search engine for SGML files in which a user can select "Right-Hand Truncation," "or search," "and search," or "not search." The CGI program gives the search condition to the OpenText search engine and receives the search result from it.

(2) Candidate functions realized with SQL

In the proposed system, the candidate function is realized with SQL Database Management System. We use the miniSQL DBMS server to handle the English Japanese dictionary data, and use the PostgreSQL DBMS server to handle with the Chinese English dictionary data. In the DBMS, there are six tables to store the historical data. Two translation tables are used to store the translation information, and four tables are used to store the metadata retrieval information. When a user consults the digital dictionary or finds metadata in the system, data in these management tables will be updated, added, deleted, or selected by the program.

(3) Displaying the multilingual fonts by MHTML

The MHTML applet is employed to display pages including multilingual characters using a browser without font. The applet is utilized here to show Chinese and Japanese characters.

5. Design of User Interface and Retrieval Method

Since the system is based on WWW, a user can access the system via Internet from all over the world. A user can input a Chinese word by Pinyin or character to get the translated English words. The user can also input two Chinese words

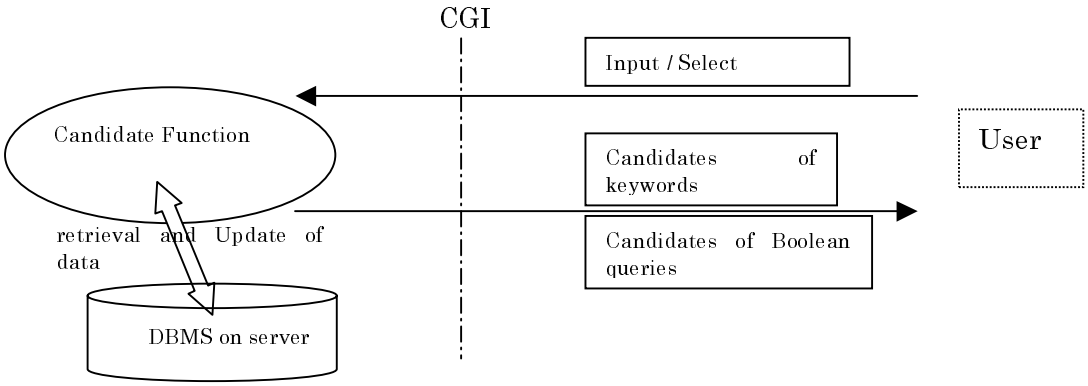


Figure 4 System Configuration

together to get two translated words. The translated words are sorted by the frequencies that they have been selected in the past.

In the case of English in Figure 5, a user inputs two English words to get two candidate lists of translated Japanese words, sorted by the frequencies that they have been selected. The lists including Japanese characters are displayed by means of MHTML technology. The user may not have the Japanese environment or may not know how to input Japanese characters. Therefore, he inputs not the Japanese words but their item numbers in front of them and also inputs a Boolean query such as "or search," "and search," or "not search."

After a user selects the Boolean query and sends it to the system, the retrieval in ULIS metadata will be started. At first, the system will return the abbreviated style.

At the same time, the system will show the candidate Boolean queries accumulated in the system, as shown in Figure 6. When a user does the retrieval with a Boolean query of "児童 and 図書館" the system will display all accumulated Boolean queries including the words "児童" and "図書館".

Examples:

(1) Boolean query "大学 not 図書館" has been selected 4 times. There are 15 retrieval results.

(2) Boolean query "図書室 not 図書館" has been selected 4 times. There are 4 retrieval results.

(3) Boolean query "大学 and 図書館" has been selected 4 times. There are 194 retrieval results.

(4) Boolean query "児童 and 図書館" has been selected 3 times. There are 10 retrieval results.

(5) Boolean query "図書館 and 公共" has been selected 2 times. There are 71 retrieval results.

The user can refer to accumulated queries and the amount of retrieval results to decide which Boolean query he will use in the next retrieval. He can select the next Boolean query from the list by indicating its associated item number in the pull-down menu at the bottom of the window.

Finally, he will be able to see the detailed style result by clicking the abbreviated style record (Figure 7).

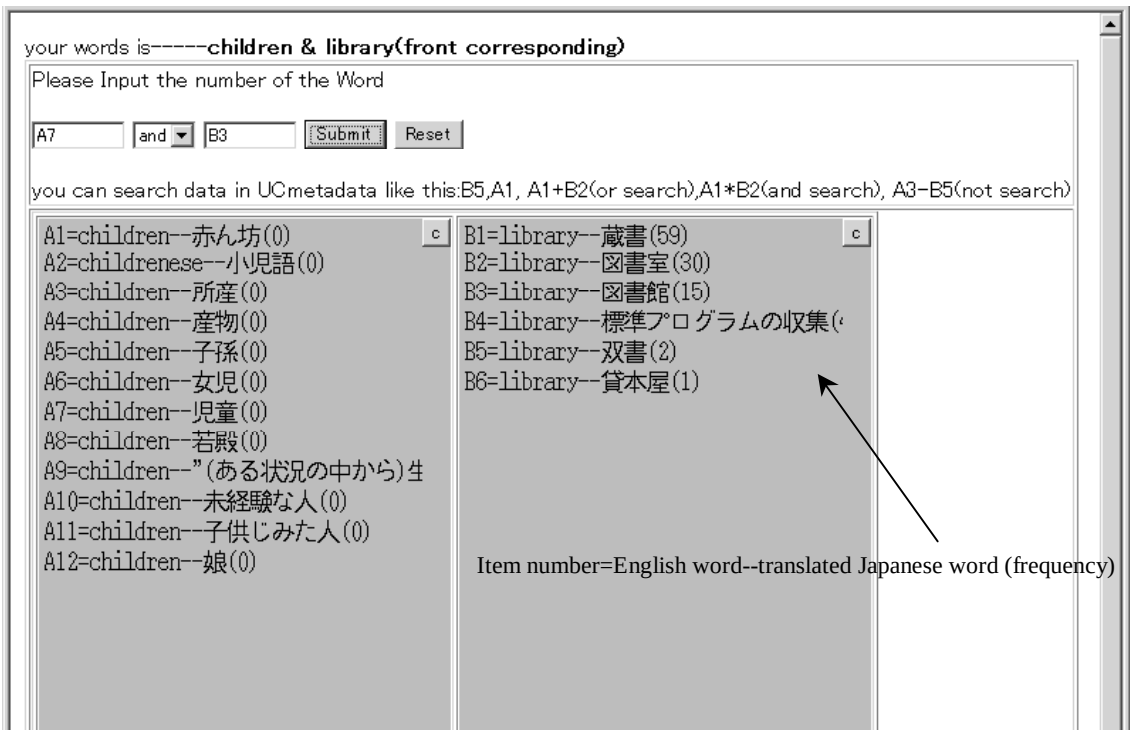


Figure 5　Example of Dictionary Result

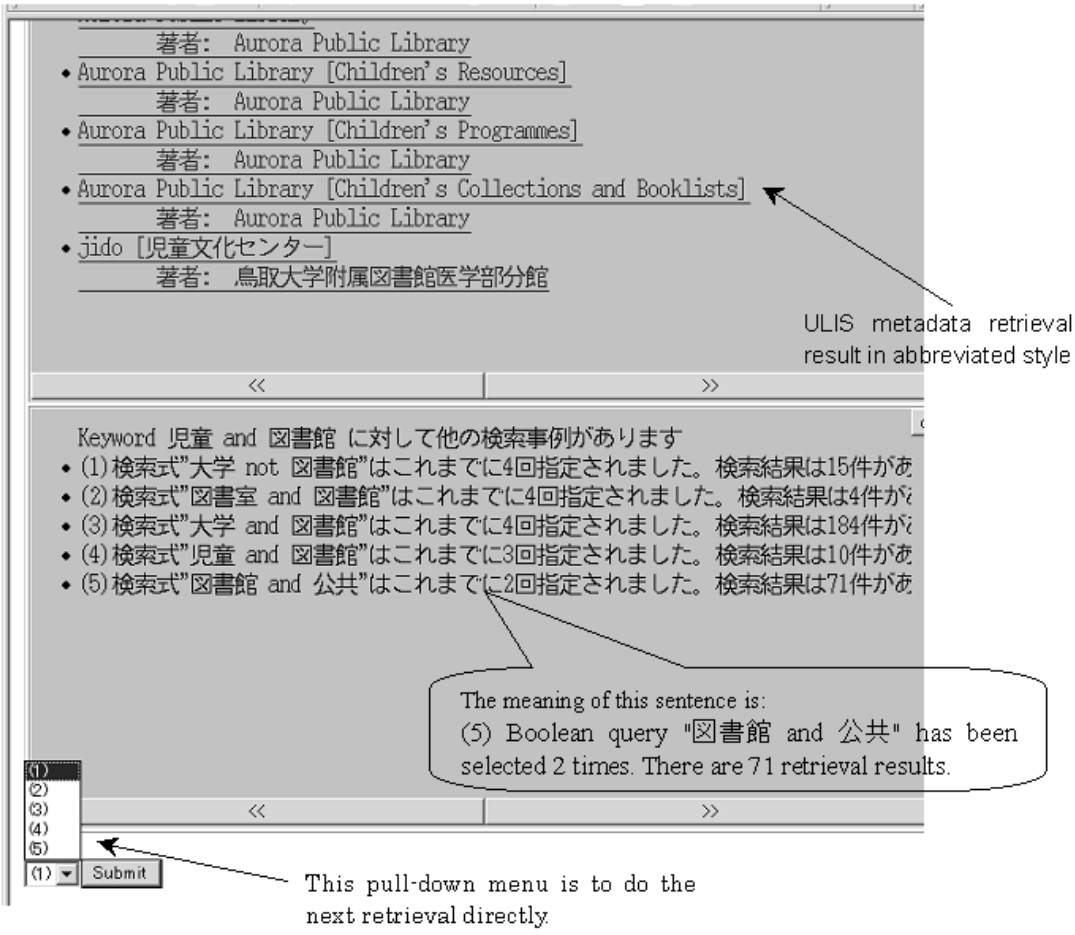


Figure 6　List of All Retrieval Results

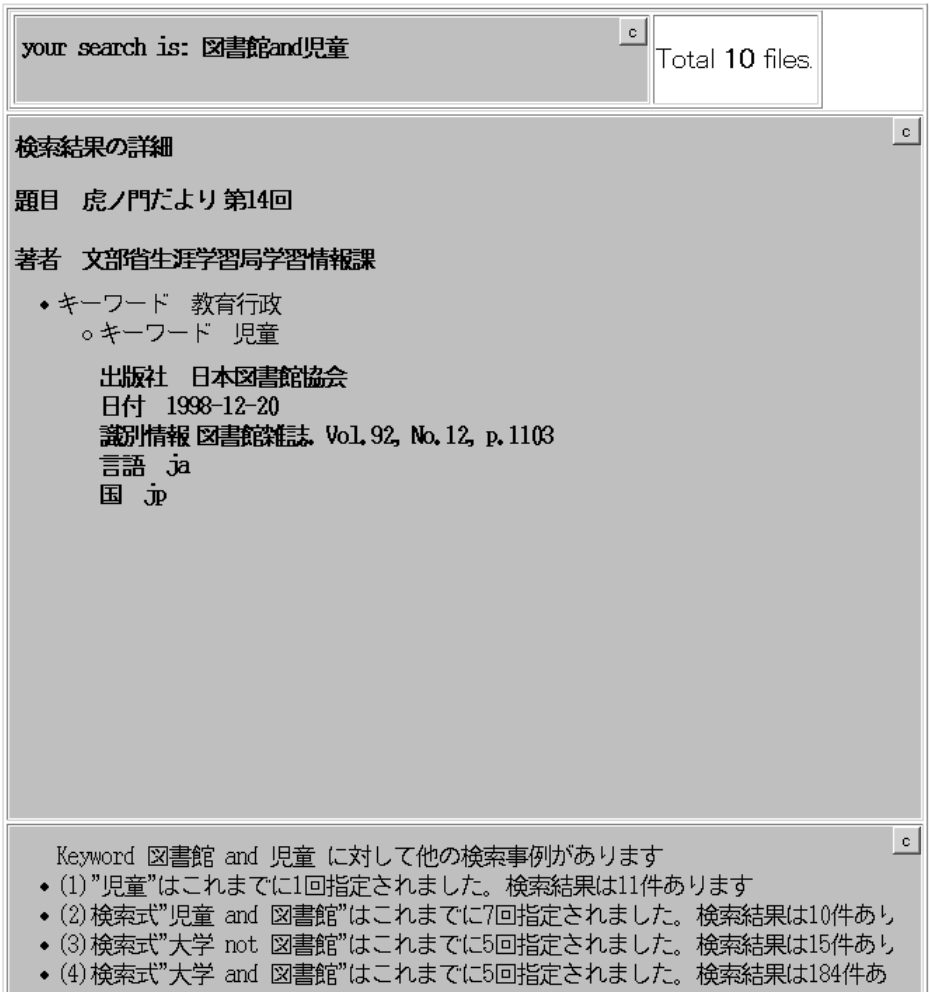


Figure 7　Retrieval Result in Details

6. Conclusions and Future Works

In the present system, users can use the system by the procedure of Chinese >> English >> Japanese >> retrieval. However, the system will soon allow access by Chinese as well. The candidate functions are useful for retrieval activities of overseas users who are not so familiar with Japanese. The system has been tested by 30 thousand ULIS metadata.

In the present system, the retrieval results are displayed by the MHTML applet. Users who do not have a Java-compliant browser cannot use the system. Therefore, the system has to provide another option to show the list of Japanese words in the image with a clickable map function on an HTML page. In the part of English and Japanese dictionary retrieval of this system, we use the DBMS of miniSQL and EUC_JP character code to handle with the words. In the part of Chinese English Dictionary retrieval, we use the DBMS of PostgreSQL and character code of Unicode. In the future, we'd like to change all Chinese and Japanese data into Unicode and be handled in only one DBMS for easily handling the database system. In this system for we are using the EDR Electronic Dictionary which can only be used within University of Library and Information science according to the copyright contract. Hence this system can’t be used by Internet outside ULIS now. On the other hand we are trying to find a free dictionary data, in order that we can make more users to use the system. We will try to integrate the ULIS metadata with the

XML data format. And we will also add other language access to the system.

Acknowledgements: The authors would like to thank Takehisa Fujita, Kyoritsu Women's University, Japan for his contribution at the early stage of this research.

References

[1] Picchi, E. and Peters, C.: Exploiting Lexical Resource and Linguistic Tools in Cross-Language Information Retrieval: the EuroSearch Approach, in First International Conference on Language Resource and Evaluation (1998).

[2] Genichiro Kikui: Retrieving Documents Across Language-Barriers — Cross-Language Information Retrieval, Journal of Japanese Society for Artificial Intelligence, Vol.15, No.4 pp. 550-558(in Japanese)

[3] Salton, G.: Automatic Processing of Foreign Language Documents, Journal of the American Society for Information Science, Vol. 21, No. 3, pp. 187-194 (1970)

[4] A. Maeda, M. Dartois, T. Fujita, T. Sakaguchi, S. Sugimoto, K. Tabata: Viewing Multilingual Documents on Your Local Web Browser, COMMUNICATION OF THE ACM, Vol. 41, No. 4, pp.64-65, April 1998

[5] Dartois, M., Maeda, A., Sakaguchi, T., Fujita, T., Sugimoto, S., Tabata, K. : A multilingual electronic text collection of folk tales for casual users using off-the-shelf browsers. D-lib Magazine. (Oct. 1997); http://www.dlib.org/dlib/october97/sugimoto/10sugimoto.html.

[6] Hiraoka H., Manaka T., Yokoyama T., Sakaguchi T., Sugimoto S., Tabata K.: Digital Library System at University of Library and information Science, Journal of Information Processing and Management, Vol.42, No.6, Sept. pp.471-479, 1999(in Japanese)

[7] http://www.dublincore.org/

[8] http://wwwj.infocom.co.jp/sgml/opentext.html

[9] EDR Electronic Dictionary: Eng.-Jpn. Bilingual Dictionary. Japan Electronic Dictionary Research Institute, Ltd. http://www.iijnet.or.jp/edr/

[10] http://www.hughes.com.au/products/msql

[11] http://www.postgresql.org/