2004－DD－46　(7)
2004／9／24

日本語 Web IME の開発と図書館情報検索システムへの実装

横山詔一*　エリク＝ロング*　米田純子*　和田志子*　黒田信二郎**　下川和男***

*国立国語研究所
**紀伊國屋書店
***イースト株式会社

あらまし　本研究は，海外のブラウザで日本語が入力できる Web IME を開発し，国内の図書館が所有する蔵書の検索システムに実装した。この Web IME は，日本語環境のないブラウザでもインターネットを介して利用が可能である。例えば，日本語をローマ字で入力し，変換ボタンをクリックすると，ひらがな，カタカナ，漢字単語などの文字列に次々と変換することができる。このような仕組みは，電子政府を支える漢字処理研究にも応用される予定であり，Unicode 外字などの入力手段の一つとして期待されている。

キーワード　IME，OPAC，国立国会図書館，電子図書館

Web IME: Web-based Japanese input method editor applied to a search system for library catalogues

Shoichi Yokoyama*　Eric Long*　Junko Yoneda*　Yukiko Wada*
Shinjiro Kuroda**　&　Kazuo Shimokawa***

* National Institute for Japanese Language
** Kinokuniya Co., Ltd.
***East Co., Ltd.

Abstract　Japanese character input mechanisms, which do not depend on operating systems and do not require installation, would be ideal for unbiased dissemination of Japanese resources. The National Institute for Japanese Language has developed a web-based Japanese input method editor, called "Web IME " hereafter, which is applied to a search system for library catalogues called "JiBOOKS". With our Web IME, users can input, edit, and display Japanese kana and kanji over the Internet without installing Japanese language systems. The databases available through JiBOOKS as of October 2004 are: 1) "Books.or.jp" of the Japan Book Publishers Association; 2) "WINE" of the Waseda University Library; and 3) English version of NDL-OPAC of the National Diet Library. Lastly, the paper addresses possible applications of and initiatives involving the Web IME at governmental levels.

Key words　Input Method Editor (IME), Online Public Access Catalogue (OPAC), National Diet Library (NDL), e-library

1. Introduction

The purpose of the paper is to introduce Web IME, an Internet-based system for inputting Japanese text, and its application to JiBOOKS, a web-based search engine that accesses data on books in print and library catalogues.

JiBOOKS was developed for two purposes: 1) to provide information on Japanese books; and 2) to display such information in authentic Japanese orthography using kana and kanji characters. First, we believe books are valuable resources that should be shared across nations, geographic locations, and languages. Second, we believe information on Japanese books should be presented in Japanese, because language is a crucial part of the art that books convey; translations may not fully represent the art of literary works.

Given these two factors, together with the mission of the National Institute for Japanese Language to enhance accessibility from overseas to information on Japan and Japanese books, we believe that there is a need for a search system with a built-in

capability to process Japanese kana and kanji. In particular, such a system provided by the Institute should be independent of technical and language settings of the users' computers so as to accommodate current needs outside of Japan, where Japanese input systems may not be available by default.

2. Database

The databases currently available through JiBOOKS are the following:

- WINE, an on-line system that contains bibliographic data on 3.4 million books among approximately 4.5 million of the entire Waseda University library collection; and
- Books.or.jp of the Japan Book Publishers Association. The data of Books.or.jp include bibliographic information on approximately 610,000 books, and is updated on a monthly basis.

In addition, as of October 2004, the National Diet Library (NDL) will link up the English version of their online catalog, NDL-OPAC, with JiBOOKS.

3. Description of JiBOOKS

JiBOOKS is a web-based search system for library catalogues. The outstanding strength of JiBOOKS is its independence from particular character encoding systems. Users do not need to install anything in order to search for and view information on Japanese books in Japanese. This is a notable advantage because overseas users, to whom Japanese language operating systems are not familiar or available, may need to install Japanese-specific systems and/or fonts in order to view Japanese on their computers, which is often a big challenge. Such installation is generally prohibited for individual users on computers on public networks, such as networks at universities and libraries, and public institutions often cannot allocate the necessary funds only to benefit a limited number of Japanese-speaking users. For home users, installation is often daunting, sometimes not even feasible, due to the lack of technical support and resources for Japanese. Further, while computers with the most recent operating systems do come with some Japanese language capabilities, there are still a great number of older computers in use, together with an increasing number of small devices with Internet capabilities such as cellular telephones. Consequently, it can be quite difficult to access digital information in Japanese from outside of Japan at present.

In order to accommodate these circumstances, a system to search and display Japanese materials over the Internet should: 1) be independent of operating systems; 2) be independent of existing character systems; 3) require no installation; and 4) be free to use. A web-based system that allows users to input and display Japanese kana and kanji characters simply by accessing a certain web site, was considered as an optimal option.

Thus, the Institute developed Web IME and JiBOOKS, which together fulfill the desired characteristics. The version currently under development may be found at the following URL:

http://btonic.est.co.jp/jibooks/78jis/

Figure 1. JiBOOKS top page

Web IME allows users to type in search terms in the Latin alphabet and convert them into kana and kanji characters. Because Web IME enables users to use kanji characters in their search terms, the target search words are better specified compared to the previous kana-only version, which produces more accurate results. For example, homophone words, such as 漱石 (name of the author Natsume Soseki) and 僧籍 (status of being a Buddhist priest) can be distinguished in the search activities, and orthographic variations found in romanization, such as Okuma / Ookuma / Oukuma / Ohkuma can be unified to 大隈.

The interface of JiBOOKS is designed to be user-friendly. The use is mostly self-explanatory, although a help page is provided. Users first access the JiBOOKS web page, then choose the database in which they wish to conduct the search. The search terms, such as parts of a book title or an author's name, are typed in romanized form and converted into Japanese characters by clicking the "Convert to Kanji" button. At first the hiragana form appears, and katakana and kanji representations are accessed by clicking up and down icons. When the target form appears, the user clicks the "Add to Search Word" button, and then "Search" to start the search. Before returning the results, the character codes for kanji and kana are converted into HTML links to GIF images, allowing users to view the results in the authentic Japanese kana and kanji orthography with any graphically based browser (Long, et al., 2003).

Figure 2. Web IME sample page 1

Figure 3. Web IME sample page 2

4. The mechanism of Web IME

4.1. Overview

This section describes the mechanism by which Web IME allows users, within the JiBOOKS application, to type in, edit, and view characters and letters, which are not supported by their browsers and computers. It requires no plug-in, only the default character input system, text, and graphic interfaces.

Responding to users' operations, JiBOOKS and Web IME work in tandem as follows:

JiBOOKS receives the user's input in the Latin alphabet together with the users' click on the "Convert to Kanji" button; it forwards this information to the Web IME service, which searches for and retrieves multiple candidate kana/kanji character strings from the data bases. These are sent back to JiBOOKS, which converts them to image data, and sends them one by one as series of image links to the user's browser. The users' browser, thus, displays each candidate conversion string in turn. Once the user accepts once of the candidates, JiBOOKS adds the corresponding kana or kanji to the search string.

4.2. Elements of Web IME

Web IME is comprised of the following elements:

1) dictionary data base for conversion;
2) Web service; and
3) IME interface embedded in a Web application.

4.2.1. Dictionary data for conversion

The dictionaries and the logic for kana-kanji conversion are installed on an Internet server. Two data bases are included in the conversion dictionary:

1.1) E1 dictionary database, and
1.2) *Daijirin* (大辞林) BTONIC Web service.

4.2.1-a. Dictionary database E1

A kana-kanji conversion dictionary called E1 is installed, which uses Microsoft SQL Server and consists of pairs of kanji and their corresponding readings. Additional words for conversion may be added to the dictionary in the CSV format. It also allows the storage of non-JIS Unicode-compatible characters.

4.2.1-b. *Daijirin* (大辞林) BTONIC Web service

This consists of the headwords of Sanseido's *Daijirin* (大辞林), which is a Japanese dictionary with some encyclopedic information. This database is searched simultaneously with the E1 dictionary previously mentioned.

4.2.2. Web service

A web service is a particular type of service available over the Internet. It is not meant to interact directly with users, but instead exchanges information with other applications that run on the Internet.

The IME web service receives strings to convert to kanji from the user's browser via the interface in the search application. More specifically, the web service provides the following functions:

1) receives the users' input in the Latin alphabet, and converts it into hiragana;
2) searches with the hiragana as keys in the conversion dictionary database;
3) merges the search results from the E1 dictionary database and *Daijirin* BTONIC (the two databases are in totally different formats); and
4) returns the search results as a series of character strings to the web application.

4.2.3. Web application

The Web IME interface is installed on the server as part of the web application that uses the Web IME service, which is the JiBOOKS search application in the paper. This application communicates with the service using the Simple Object Access Protocol (SOAP), which is a variety of XML commonly used to communicate between applications and services over the internet.

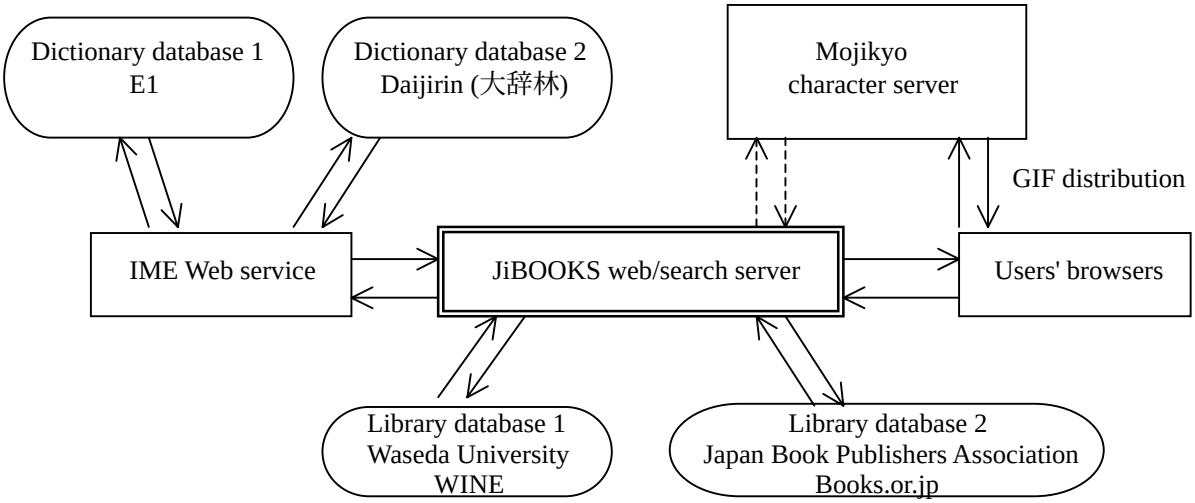

Figure 4. Description of Web IME and JiBOOKS

The web application mediates communication between the users' browsers and the web service and between the users' browsers and the Mojikyo character server.

In specific, the web application provides the following functions:

1) relays the search words which users have input in the Latin alphabet, to the web service by using SOAP;

2) looks up the kanji and other characters returned from the web service, to the Mojikyo font server; and

3) provides graphic representations of the characters in the form of links to GIF files on the Mojikyo server.

4.3. Applications of Web IME

The mechanism of Web IME, which converts the Latin alphabet to orthographic representations in other languages and then to image, consequently allows flexible applicability. For example, if Chinese is included in the dictionary and the font servers, it can convert Pinyin into Chinese characters. Likewise, it can provide characters not included in JIS (the characters defined by the Japanese Industrial Standard), as long as the pronunciation is available. In other words, additional functions can be easily installed only by adding data to the dictionaries and the font server, depending on the given contexts. Such flexibility is, in fact, a generic characteristic of Web services.

Furthermore, since Web IME operates over the internet with SOAP, a simple protocol that is implemented on almost all operating systems, Web IME can be easily installed on server applications on Windows, Unix, Mac and other machines.

5. Application of JiBOOKS technology

The primary purpose of the JiBOOKS project was initially two-fold: 1) to promote accessibility to Japanese information from overseas; and 2) to assist libraries and library users. The project has produced positive outcomes. In fact, an evaluation survey study conducted in Malaysia strongly supported the contribution of JiBOOKS from users' points of views (Yokoyama et al., 2004). Furthermore, the decision of the National Diet Library (NDL) to employ our system also suggests the significance of JiBOOKS in library science.

The project was initially targeted to libraries and library users, however, it should be noted that the project is based on the achievement of a collaborative research project. The development of Web IME was partially supported by the e-Japan initiative of the Institute, a plan to promote the use of Information Technology in support of teaching Japanese as a second language. More significantly, the project is connected with a database project to organize and make available all kanji which are needed to express Japanese personal and place names in government documents.

The research team for the kanji database project includes the National Institute for Japanese Language (国立国語研究所), Information Processing Society of Japan (情報処理学会), and the Japanese Standards Association (日本規格協会). The project is sponsored by a group of five governmental agencies and ministries, such as the Ministry of

Economy, Trade and Industry (経済産業省).

Although the major purpose of the kanji database project is to construct a database, it also aims at standardization of kanji for electronic government. In developing the kanji database, JiBOOKS, and other digital tools within the framework of the electronic government initiatives, several criteria were set as to what characters and their various forms should be made available in the system. For example, it was decided to include external characters, i.e. gai-ji (外字) which are not encoded in the JIS standard, by representing them in the GIF format. Examples of gai-ji that were judged as critical items include the characters used in names of authors, such as the "ou (鴎)" as in Ougai Mori (森鴎外). Such a decision was necessary because names are conventionally represented by the authentic forms in printed materials, including textbooks. Thus, some gai-ji might be more familiar and popular than JIS characters. In fact, studies have shown that native speakers of Japanese are more familiar to some gai-ji than their counterpart revised JIS forms. Therefore, we concluded that our system should be able to handle gai-ji, which are frequently used in the printing industry.

Another strength of our system to be mentioned is that our Web IME has potentials for use in a variety of applications. For example, it may be used in on-line materials for teaching Japanese as a second language. In fact, Web IME is implemented in JiWODRS, which is a system of on-line dictionaries (http://btonic.est.c.jp/JiDic/). JiWORDS is at this point a pilot project in a trial phase. The dictionaries available through JiWORDS are *Daijirin* (大辞林) Japanese monolingual dictionary, *Daily Concise* English- Japanese and Japanese- English dictionaries of Sanseido (三省堂). At the governmental levels, the National Printing Bureau (国立印刷局) is considering applying our Web IME technology to the overseas distribution of *The Official Gazette* (官報) over the Internet. Likewise, our Web IME may be applied to digitalization of characters that JIS cannot do so.

The movement to accommodate gai-ji needs is found in the industry as well. An example is a commercially available product called "Interstage Charset Manager" by Fujitsu (2004). It enables gaiji. input through the interface that is familiar to the users as it resembles to popular IMEs. Given that such a product has been commercially successful and that kanji is a critical factor in documents in Japanese, we believe that the gaiji issues continue to be a major and popular interest in the field.

References

Fujitsu (2004)
http://interstage.fujitsu.com/jp/output/charsetmgr/index.html

Long, E., Yokoyama, S., Kumagai, Y., Yoneda, J., & Kess, J. F. (2003)
JiBOOKS: An Image-based Japanese-language Data Retrieval System. In *"Changing Japanese Identities in Multicultural Canada" Conference.* Centre for Asia-Pacific Initiatives 2003, University of Victoria, Canada. pp. 465-471.

Yokoyama, S., Lee, S. L., & Ishida, T. (2004)
Bibliographic catalogue web-based search system designed for non-Japanese browsers "JiBOOKS": Report on evaluation survey in Malaysia. National Institute for Japanese Language, Japan.