Missing Character Management in Digitalization of the Hankuk Pulgyo Chonso

Keum Suk Lee, Young Sik Hong, Yong Kyu Lee
Computer Engineering Dept. and Electronic Buddhist Text Institute (EBTI)

Tae Sik Han (Ven. Bo Kwang Sunim)
Seon Studies Dept. and Electronic Buddhist Text Institute (EBTI)
Dongguk University, Republic of Korea

Abstract

We have developed a new text editor based on the Unicode to digitalize the Hankuk Pulgyo Chonso (Korean Ancient Buddhist Corpus), which is a complete compilation of Korean Buddhist Works. Even though the Unicode supports more than 27,000 CJK characters, we have found many missing CJK characters that cannot be inputted with the Unicode.

So, we have extended the text editor's functionality in order to input and handle missing characters with a menu-driven style. With the text editor, we can register a missing character and map into a corresponding CJK font. The missing character fonts have been downloaded from the Mojikyo Institute (http://www.mojikyo.gr.jp) and stored in a local database. Also missing characters can be displayed with web browsers such as Netscape Communicator and Internet Explorer.

The text editor and database may be useful for studying missing character types and also for digitalizing other ancient Buddhist texts of Korea. We are going to analyze the statistics of the missing characters in the Hankuk Pulgyo Chonso.

1. Introduction

Most Korean ancient documents have been written in CJK characters and several institutions, such as the Research Institute for Tripitaka Koreana, the Seoul Systems Inc. and the Electronic Buddhist Text Institute of Dongguk University are building their own full-text databases of Korean ancient documents. Thus, it is necessary to use an appropriate text editor to key in ancient source documents and build the full-text databases. However, most Korean text editors based on KSC-5601 do not support more than 4,888 CJK characters. Even though some editors based on 4-byte code can handle more than 30,000

CJK characters, web browsers cannot display such 4-byte code characters due to the incompatible code system.

So, we have developed a new text editor based on the Unicode with SGML-based markup functions to solve these problems and to help build Korean ancient full-text databases [7]. And our text editor is to be improved to handle XML-tags and used to input parts of the Hankuk Pulgyo Chonso (Korean ancient Buddhist corpus) which has already been on the web server and may be accessed through the Internet now (http://ebti.dongguk.ac.kr).

Even though the Unicode supports more than 27,000 CJK characters, we have found many missing CJK characters that cannot be inputted with the Unicode. In this paper, we describe the process of missing character management in digitalization of the Hankuk Pulgyo Chonso using the text editor. The Mojikyo fonts [9] from the Mojikyo Institute are used to solve the font problem of missing character. Missing character management system using the Mojikyo fonts and the method for uploading the fonts on the web are described on the next chapters.

2. Text Editor based on the Unicode

It is very important to use an efficient text editor for processing Korean ancient documents written in CJK characters. In Korea, there are several commercialized text editors, which can handle only 4,888 CJK characters based on the KSC (Korean Standard Code Set)-5601. Even though some text editors can handle more than 30,000 CJK characters, CJK characters processed by these text editors cannot be browsed with the existing WWW browsers, such as Netscape Communicator and Internet Explorer due to incompatible codes.

We have designed a new text editor to help build a Korean ancient Buddhist full-text database. Our text editor supports all the CJK characters of the Unicode and has SGML-based markup and error checking functions. Figure.1 shows the main window of the editor.

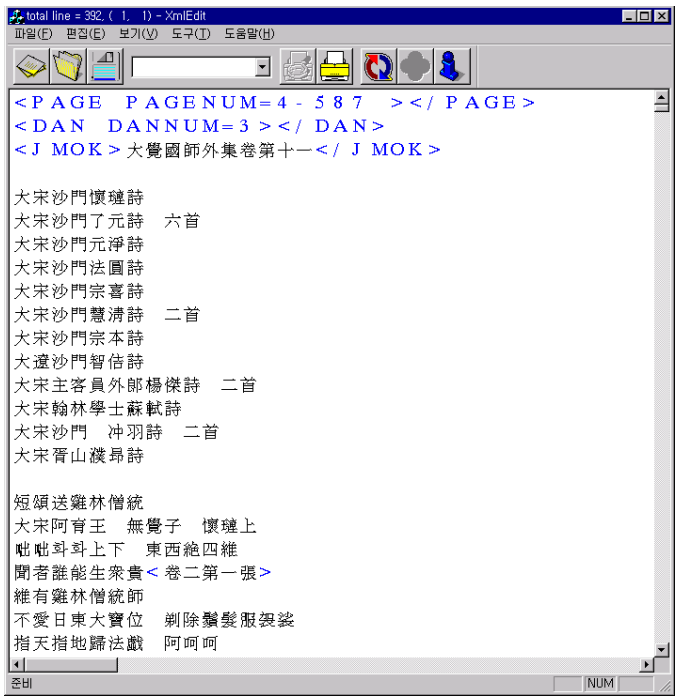

Figure.1 The text editor using SGML and Unicode

The text editor supports all the CJK characters of the Unicode. However, there are many missing characters on the Korean ancient documents, which are not supported by the Unicode.

First of all, we need font images to handle these missing characters. One solution is to design font images for missing characters. The other solution is to use previously published font images from other organizations. We choose the latter and we used the Mojikyo fonts from the Mojikyo Institute.

The Mojikyo fonts not only include all the Unicode CJK characters, but they also support a lot of special CJK characters, which totally amount to about 90,000 characters. To support loading these fonts on the web documents, the Mojikyo Institute have already stored the fonts in the form of GIF image on its web server, and supports user to generate an URL to specify the location of the selected font [9]. Users only need to insert the generated URL to their HTML documents.

Figure.2 shows the Mojikyo font manager. One can click one of the traditional radical component of a CJK character on the left side of the window and then the right side of the window shows the CJK characters include that radical. If one clicks on the selected character, the menu that supports to generate an URL to the selected character is displayed.

Figure.2 The Mojikyo character map

An example of the generated URL of Mojikyo font inserted in a HTML document is shown on Figure.3. Figure.4 shows the result on the web browser for the HTML document.

```
<HTML>
    <HEAD>
        <TITLE>Test of the Mojikyo Font Image </TITLE>
    </HEAD>
    <BODY><B>
        A Chinese character in the Mojikyo fonts can be displayed on the Web page by
        using the URL to the font image stored at Mojikyo
        homepage(http://www.mojikyo.gr.jp).<BR>
          <hr>
          <p align=center>

        <IMG WIDTH=30 HEIGHT=30 NAME="mojikyo_font_014457"
```

Figure.3 HTML source to insert a Mojikyo font on web page

Figure.4 The result window showing a Mojikyo font

However, the above processing of the missing character is strongly dependent on the Mojikyo site. It is impossible to mange the used fonts for our own demand. So, we proposed to transform the Mojikyo fonts to images and use the images to construct our missing character manager and database.

3. Management of Missing Characters

We propose a scenario which represents the steps to input the Korean ancient documents including missing characters. When a missing character is found, the user searches the character using the Mojikyo font manager. And then, the font is copied to the clipboard on the Windows operating system. The copied font can be stored as the BMP format on one's own local disk or can also be transformed to other graphic format, such as GIF and JPG. We transformed the font for missing character to the GIF format and registered the font in our missing character database with other important information, such as the radicals, the stroke count, and Korean pronunciation.

Figure.5 shows the entire process of the digitalization of the Hankuk Pulgyo Chonso including the process for the management of missing characters.

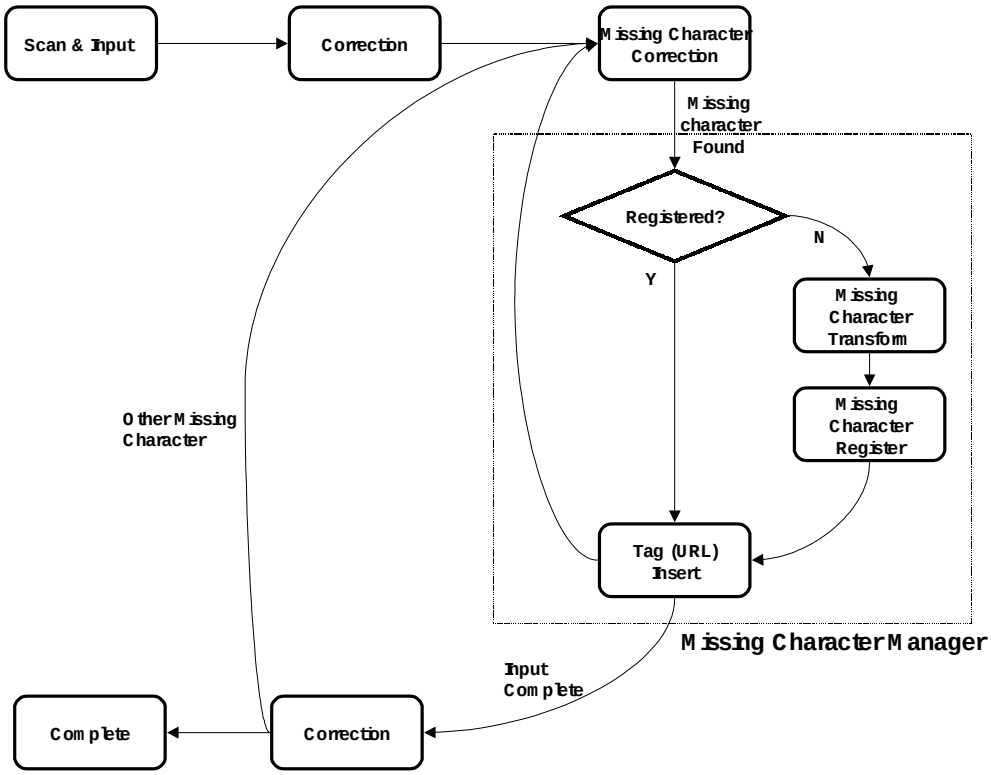

Figure.5 Process diagram for digitalization of Korean ancient document

When the user finds a missing character, he can search the found character in our missing character database. If the character is not registered on the database yet, he can register the character by using missing character manager through the described steps above.

Figure.6 shows the main window of the missing character manager. We used Microsoft Visual Basic 6.0 for developing the missing character manager on the Windows NT operating system. Microsoft Access 7.0 is used for DBMS.

Figure.6 The main window of missing character manager

The window for registering a new missing character to the database is shown on Figure.7. To register a missing character, information regarded with the character, such as the radicals, the stroke count, and Korean pronunciation are stored in the database. Newly registered missing character becomes to have a unique code, which is used as the file name for the font image.

In the meanwhile, the registered missing character is inserted on the documents in the form of URL like the URL part of Figure.7.

Figure.7 The window for registering a missing character

Figure.8 shows the search window of the missing character manager. A missing character can be searched by the stroke count of the found character.

Figure.8 The window for searching a missing character

Registered missing character is automatically inserted in the form of URL on the document during the registration or search process.

Figure.9 represents the generated URL by the missing character manager, and Figure.10 shows the result web browser window.

As one can see in the HTML source, the registered fonts are represented in the <IMG> HTML tag.

```
<HTML>
    <HEAD>
        <TITLE>Test of the Mojikyo Font Image </TITLE>
    </HEAD>
    <BODY><B>
        A Chinese character in the Mojikyo fonts can be displayed on the Web page by
        using the URL to the font image stored in our missing character database.<BR>
          <hr>
          <p align=center>
```

Figure.9 HTML source to insert a Mojikyo font on web page

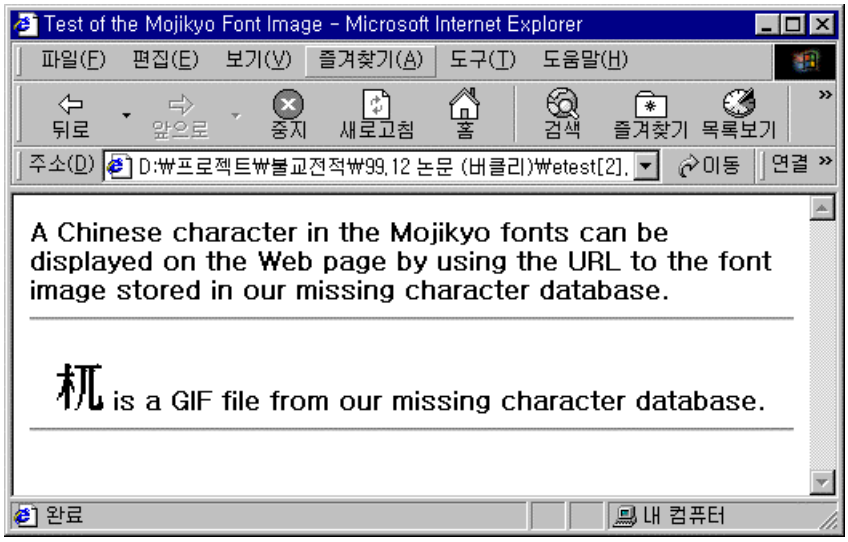

Figure.10 The result window showing a Mojikyo font using our database

4. Conclusions

In this paper, we have described our text editor based on the Unicode to digitalize the Hankuk Pulgyo Chonso (Korean Ancient Buddhist Corpus). Even though the Unicode

supports more than 27,000 CJK characters, we have found many missing CJK characters that cannot be inputted with the Unicode.

So, We have extended the text editor's functionality in order to input and handle missing characters with a menu-driven style. With the text editor, we can register a missing character and map into a corresponding CJK font. The missing character fonts have been downloaded from the Mojikyo Institute and stored in a local database. Also missing characters can be displayed with web browsers, such as Netscape Communicator and Internet Explorer.

The text editor and database may be useful for studying missing character types and also for digitalizing other ancient Buddhist texts of Korea. We are going to analyze the statistics of the missing characters in the Hankuk Pulgyo Chonso.

[References]

[1] Dongguk University Press, The Hankuk Pulgyo Chonso (Korean ancient Buddhist corpus), Vol.1-12, 1979-1998. (In CJK)

[2] Dongguk University Press, The Hangul Tripitaka, Vol.1-280, 1996-1998. (In Korean)

[3] C.F.Goldfarb, The SGML Handbook, Oxford University Press, 1990.

[4] Y.S.Hong, et al., "Development of the Technologies for Korean Ancient Document Management and Retrieval on the Web," Project Final Report, Ministry of Information and Communications, 1998.

[5] Chu-Ren Huang, Keh-Jiann Chen and Shin Lin, "Corpus on Web: Introducing the First Tagged and Balanced Chinese Corpus," PNC Special Meeting in Taipei, Feb.17-19, 1997.

[6] The Unicode Consortium, The Unicode Standard, Version 2.0, Addison-Wesley, 1996.

[7] Y.S.Hong, et al., "Development of a Syntax-directed SGML Editor for Processing Korean Ancient Documents," Proceedings of 1999 EBTI, ECAI, SEER & PNC Joint Meeting in Taipei, Jan. 18-21, 1999, pp.143-149.

[8] Y.K.Lee, et al., "The Hankuk Pulgyo Chonso and the Hangul Tripitaka (the Korean Ancient Buddhist Corpus and the Korean Translation of the Koryo Buddhist Canon) on the WWW," Proceedings of 1999 EBTI, ECAI, SEER & PNC Joint Meeting in Taipei, Jan. 18-21, 1999, pp.395-406.

[9] Tokio Furuya and Tsuneo Yatagai, "The Activities of the Mojikyo Font Center," Proceddings of 1999 EBTI, ECAI, SEER & PNC Joint Meeting in Taipei, Jan. 18-21, 1999, pp.329-337.