On the Missing-Characters (Gaiji) of the Taisho Tripitaka Text Database Published by SAT

Shigeki Moro
The Association for the Computerization of Buddhist Texts, Japan

0 ABSTRACT

In March of 1998, the Association for the Computerization of Buddhist Texts (ACBUT) began publishing the electronic text database of the Taisho Tripitaka. SAT is the nickname of this project.

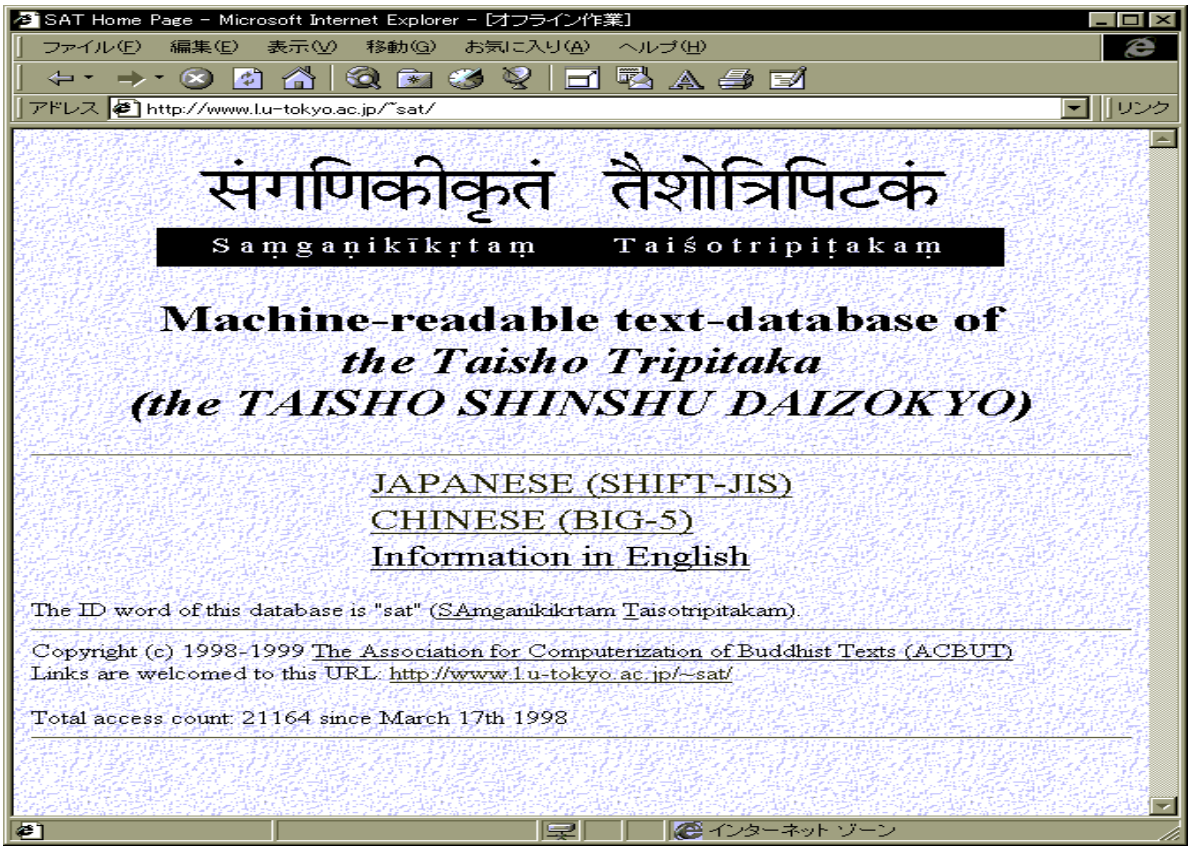
The Taisho Tripitaka includes both classical Chinese and Japanese texts, so that SAT texts are encoded by JIS code set at the present. In the not-too-distant future, they shall be changed to larger sets like Unicode. But there always are characters that can not be input. The solution of the Gaiji (missing characters) are the most important subject for the projects like SAT. Now SAT has about 90 published e-texts, and they include over 7 million characters. Over 17,000 characters cannot be input with JIS and about 1,500 with Unicode.

Following the KanjiBase developed by Dr. Christian Wittern, we now use SGML-style placeholders that are both standardized and system-independent. And we are investigating the empty-element-tag of XML as a new solution.

1 What is SAT?

1.1 Outline

SAT is the nickname of 'Samganikikrtam Taisotripitakam" in Sanskrit. Aiming to publish all the electronic texts of the Taisho Tripitaka (**大正新脩大藏經**), the Association for the Computerization of Buddhist Texts (ACBUT) began to develop SAT in 1994, and the Chinese translation of the Maha-prajnaparamita-sutra was published on Internet in March of 1998. Now about 90 texts are available. Anyone can download them from the our web site (http://www.l.u-tokyo.ac.jp/~sat/, Photo 1). They are freeware and no commercial use is allowed in principal.



[Photo 1: Homepage of SAT]

Most texts of them are input with OCR programs, and some texts are by hand. I would like to emphasize the cooperation with the Chinese Buddhist Electronic Text Association (CBETA) for inputting, proofreading or so forth. In the near future, along with the Text Encoding Initiative (TEI), CBETA and SAT will use common minimal markup such as Metadata or headers to provide information about the file and its sources.

1.2 Technical Information

The Taisho Tripitaka includes Chinese characters, Japanese Hiragana and Katakana, and rare characters such as Siddham. So that SAT texts are encoded by Shift-JIS (JIS X 0208:1997) and SGML-style placeholders at the present. In the not-too-distant future, they shall be changed to UTF-8 or UTF-16 of Unicode which will be the standard code set of XML and Internet. But there always are characters that can not be input. As a solution for the problem of the missing characters (Gaiji in Japanese), SGML-style placeholders are used; but we will get to that later, after we see some tables on the characters of SAT.

Those texts available on our web site are formatted with some basic tags mainly for the text retrieval system such as grep. Various formats will be provided to suit the needs of different

uses. For example, now published texts don't include the information of the footnotes and Kaeriten. Kaeriten is a group of symbols for Japanese to help reading classical Chinese texts easily. Needless to say, the footnotes and Kaeriten are useful information in order to study Buddhist texts. And we have input both of them. So that we will be able to provide a version with the footnotes or Keriten.

2 Characters of SAT

The first question which the database developers of the oriental studies encounter is, how many Gaijis do we need? How many characters will the database use? Of course, the correct answers are given after we finish the projects, but it may be helpful for developers to get the data of the characters in our databases on hand.

The following data were collected from the bodies of the 88 electronic texts published by SAT, not from the footnotes of them. The 88 texts include 71 translations of Indian texts, 4 Chinese, 2 Korean and 11 Japanese texts. Most texts are written in classical Chinese, but some texts are written in classical Japanese.

Table 1 shows the number of characters of SAT. Over 7 million characters in totality include 17278 characters that cannot be input with JIS code set, 1579 lacks with Unicode, and 428 that cannot be found in the Morohashi Chinese-Japanese character dictionary. The percentage of this data seems very small. However, since roughly 0.06% means one missing character per page of Taisho Tripitaka, it seems to me that this result is not small but serious.

Table 1: **Number of characters**

	ALL (%)	IND (%)	CHN (%)	KOR (%)	JPN (%)
Total	7,075,903 (100)	5,954,186 (100)	724,142 (100)	55,428 (100)	342,147 (100)
Not in JIS	17,278 (0.244)	15,317 (0.257)	921 (0.127)	27 (0.049)	1,013 (0.296)
Not in Unicode	1,579 (0.022)	1,411 (0.024)	72 (0.010)	5 (0.009)	91 (0.026)
Not in Morohashi Dic.	428 (0.006)	375 (0.006)	23 (0.003)	1 (0.002)	29 (0.008)

Table 2 shows the kinds of characters of SAT. At the present SAT is made up of 5,480 varieties of character. Of course, as the number of text increases, the number of character also increases, but I think that less than 15,000 characters are adequate to input the whole of the Buddhist canon written in classical Chinese and Japanese. It is said that Unicode will add a new Chinese character set called "Super CJK" which may be equal to the Emperor Kangxi's **康熙** dictionary. But the addition is not enough for the Buddhist database.

Table 2: **Kinds of characters**

	ALL (%)	IND (%)	CHN (%)	KOR (%)	JPN (%)
Total	5,480 (100)	4,747 (100)	2,767 (100)	1,143 (100)	3,252 (100)
Not in JIS	1,264 (23.06)	948 (19.97)	184 (6.65)	22 (1.92)	375 (11.53)
Not in Unicode	338 (6.17)	283 (5.96)	30 (1.08)	4 (0.35)	48 (1.48)
Not in Morohashi Dic.	160 (2.92)	136 (2.86)	11 (0.40)	1 (0.09)	19 (0.58)

As Dr. Chuang Teming of Academia Sinica points out, over 99% of characters in the 25 dynastic histories are made up of about 5,000 kinds of Chinese character. He says:

以二十五史為例，二十五史的總字頻次為 32,479,141，總字數為 13,955，其中使用頻度最高的 5,000 字的總字頻次就佔了 99.57%[1]

Although the conditions are slightly different, the result on the 25 histories is similar to the SAT's (Table 3). Almost the whole part of SAT is made up of the basic characters (Table 4), and the frequency of Gaiji is very small.

Table 3

If we use...	... could be made up
Top 10 kinds	17.05
Top 100	63.9
Top 1,000	98.5
Top 5,000	99.9

Table 4: **Top 10 characters**

無	(221,780)
不	(151,830)
若	(135,533)
故	(122,259)
淨	(109,969)
一	(96,900)
是	(95,771)
薩	(94,108)
有	(88,143)
清	(87,653)

[1] **莊德明「漢字缺字處理與梵巴藏字母的輸入」『佛教圖書館館訊』第 14 期、1998 年 6 月**

3 Gaiji Solution of SAT

3.1 Entity

In order to develop a useful and reliable database, a method for proper handling of Gaiji has to be found. There are some solutions to handle as shown below.

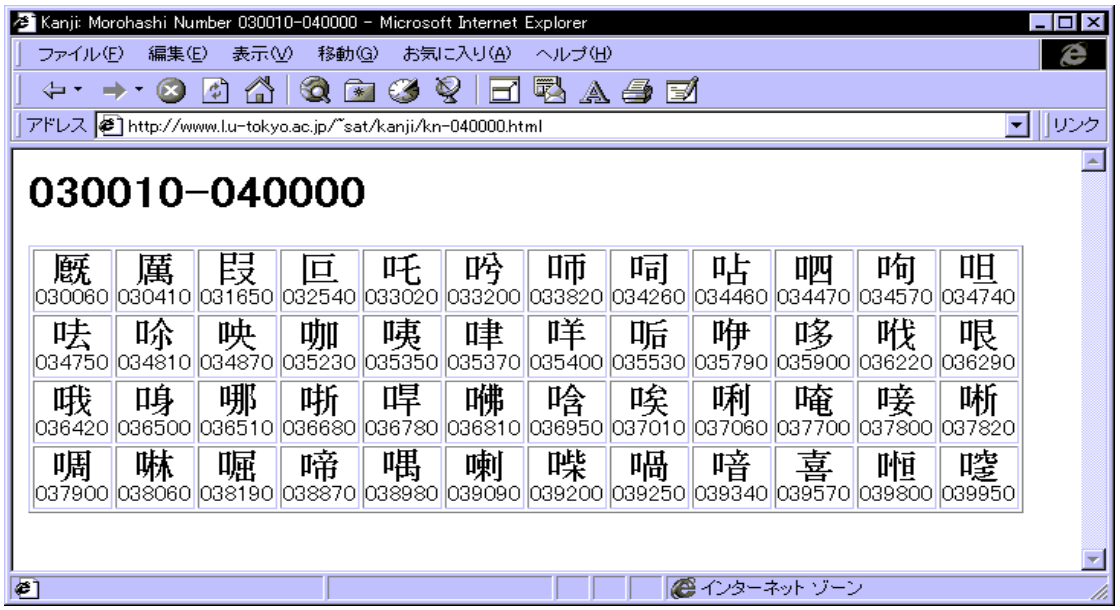
* Composing characters from components available in the system
* Using SGML-style entity references (either KanjiBase codes or Mojikyo numbering)
* Using the composition method such as developed at Academia Sinica

At the present the Gaiji of SAT are handled with SGML-style placeholders following the KajiBase developed by Dr. Christian Wittern. For example:

&M-123450;

Ampersand and semicolon are the opening and closing delimiters. The following "M" signals that the following is a code of the Morohashi dictionary or that of the Konjaku Mojikyo. After the dash, the left 5 figures mean the code. In the range from 1 to 49,964 are the Morohashi codes, and beginning at 49,965 are the Mojikyo codes. The right 1 figure is in reserve.

Users who have some difficulties to access the Morohashi or Mojikyo can make sure of the forms of JIS Gaiji on web (Photo 2).



[Photo 2: Gaiji reference page on WWW]

3.2 Empty element tag of XML

If we suppose that a new common DTD is developed for the Buddhist canon databases such as SAT and CBETA, all entities would be declared in it. But, as has been noted above, there are too many kinds of Gaiji to be handled by SGML/XML parser. Nevertheless the frequency of Gaiji is very rare. So that I suggest to adopt to the empty element tag of XML as a new solution. For example:

<外字 charset="文字鏡" n="12345"/>

Certainly it is a demerit that the placeholders of the empty element tag are too long, but it is the adoptable merits that the DTD could become simpler and easier to maintain, because only one structure of the tags must be declared in DTD. And in using XML we will be able to use flexible hyper-linking functions such as Xlink, XPointer and Namespace. And we will be able to browse the texts with free and popular browsers such as Internet Explorer version 5, or Netscape Communicator version 5.