The SAT Project and Web Solution

Nagasaki Kiyonori*

목 차

1. Introduction

1.1. The prologue of the SAT project

The Japanese Association of Indian and Buddhist Studies (JAIBS) has been engaged in digitizing their materials since the early days of personal computer technology in order to assist the research by its members and others who have been interested in such studies. In 1986, their first results were published. It was a report that suggested some solutions through the building of the bibliographical database of their journals. In 1989, JAIBS established a database center. Since then, JAIBS also has tried to digitize various materials to assist

* International Institute for Digital Humanities, Japan

members and others.

1.2. Complete transcription of Taisho-Daizokyo

The late professor Yasunori Ejima aimed for transcription of the entire text of Taisho-Daizokyo, including around 6 million lines in 85 volumes composed of Buddhist canons compiled in India, China and Japan, into electronic text data so that any researcher interested in it could search and process it easily on their computers. In 1994, he started the project to input some text data into computers, while he was funded by Grant-in-aid for scientific research of The Ministry of Education, Culture, Sports, Science and Technology, Japan (MEXT Grant). And then, The Association for Computerization of Buddhist Texts (ACBUT) was established to manage the project and the database project was named "Saṃgaṇikīkṛtaṃ Taiśotripiṭakam" (SAT). (The association is directed at present by Masahiro SHIMODA.)

It was quite difficult to transcribe the entire text onto computers. The first and most important problem was exact transcription. It might have been impossible because of the large size of Taisho-Daizokyo. The digitized data could only be made as close to the original text as possible. With acceptance of this point alone, this project could be considered significant.

1.2.1. Validation of the format

The text data has a format that allows representation of various annotations and other information. At least, each line must have its own location such as text number, volume number, page number, paragraph sign and line number.

Moreover, some XML-like tags are prepared in order to represent various types of information. It was a good suggestion, but difficult to describe correctly because the format was not so stable, the validation of the format was not prepared, and the operation describing it by using such tags was not so easy.

1.2.2. "Gaiji" problem

The old Chinese characters used in Taisho-Daizokyo are of great variety. Though most of

them are defined in the character encoding standards such as Unicode, some characters are not. It is difficult to decide whether the subtle difference between similar glyphs of the characters means the difference of the characters, the glyphs or only the mistake in the process of the publication in Taisho-Daizokyo because it sometimes depends on the context of each book rather than the nature of each character or each glyph. At first, in order to represent those characters, we adopted a large font set, but changed into another font set because of a license restriction as described later. Moreover, we had to prepare an original character code set and each glyph image which do not exist even in the font set, because we had to represent those exceptional characters, which are called "gaiji," in the guideline of the TEI (Text Encoding Initiative) and in Japanese. Such a problem was solved by the "gaiji" database to be described later.

1.2.3. Differences in computer skills of each collaborator

This project was not developed by computer engineers or keypunch operators, but done almost entirely by researchers of Buddhism. This policy is basically good to build such database from the point that demands of the users are reflected directly and easily. However, their computer skills are all different and parts of the data were not exact.

1.2.4. Checking the status of each process by each collaborator

In this phase, collaborators exchanged their active text data directly with floppy disks or other removable digital media. In such a state, it was difficult for each collaborator to carry out the process.

1.3. Transcriptions were finished but had many errors in characters and formats.

In 2005, the first phase of the project was finished. The entire text was digitized and released on its Web site. However, some problems remained.

One of them was the typing errors in the text body. In order to correct them, the project

had Buddhist researchers check text data character-by-character comparing them with the original texts of the first edition of the Taishō Daizōkyō, and with the database texts released by CBETA in 2005.

The other problem became explicit when the project ordered some viewer software from a company which could represent their text data just like the layout of the original book. It was necessary to correct all of the mistakes in the format of the text data. It was difficult to keep in touch with each collaborator to train and upgrade their computer skills and unify the format in detail because they lived in various separated places in Japan. To solve the problem, the project began to develop a Web collaboration system which they could operate easily and exactly in October, 2005. The project started using the system in January, 2006.

2. A Web collaboration system for the purpose of correcting text data

2.1 Correcting text through Web

As described above, the project started using a Web collaboration system which enabled each collaborator to correct the text data from their own place anytime and its operators to manage those corrections. Collaborators needed to pass authentication in order to access or edit the text data. Then, they could choose a title page of a text for their own task. The title page provided some functions:

a. to download each book into a local computer as a text file.
b. to display each page of the text including “gaiji” characters on their Web browser.
c. to rewrite each line on their Web browser.
d. to upload each book from a local computer.
e. to search each line by keywords.
f. to refer lines which were corrected previously and the working log including the name of

the collaborator and the time.

If a collaborator corrects a line, the name, the time and the previous line are recorded in the system. If a collaborator uploads a text file, the system validates the format of the text file and corrects lines which are different from the old by comparing with those in the system.

The system provides two methods in order to correct the text. One is to rewrite directly on the Web browser, and the other is to edit on each local computer and to upload the text to the system because it assumed there were two types needed by collaborators. Beginners use the former and specialists needed the latter. At this time, while active interfaces for the Web browser become more common such as internet shopping or internet banking, the average PC users also have become familiar with such interfaces. This movement was called "Web 2.0". However, some specialists built their own software environments on their local computers to correct text data efficiently. Therefore, to keep their work efficiently, it was necessary to be able to work on such environments continuously and to have use of the latter method.

2.2. "Gaiji" database

At this time, the project adopted one of the large font sets, called GT-font[1], which could represent over 60,000 characters or glyphs. However, some characters were not in the font sets. Unrepresented characters were inserted into the "gaiji" database with some information about each character included in the system. Some designers drew glyph images to represent the "gaiji".

3. Two products of the project at this time[2]

1) See http://www.l.u-tokyo.ac.jp/GT/ .

As described above, in July 2007, over 50 researchers have basically finished the task of correcting the text data. Moreover, the project made vector image files of the Indian characters, around ten thousand in all. To publish the results of the project, two products were prepared. One was viewer software which could display each page just like the original book and search the text data. The other was the search system on a Web site which could rapidly search and distribute the entire text data. The two will be explained below.

3.1. The viewer software

The viewer software can run on various local computers such as MS-Windows, Mac OS, Linux and so on because it is implemented by Shockwave Flash. Reading XML-like tags marked up on the text data, it represents each page just like the original book. However, it has a few minor problems, for example, search speed, stability or extensibility.

3.2. The Web site to search the entire text and more

To solve the problems of the viewer software and to offer more services for those who are interested in Buddhism, the project is now developing a new Web site using free software such as Linux, Apache, PHP and PostgreSQL to search the entire text data rapidly and to produce the text data in XML format. One of the functions is related to character search by use of a character ontology data built in XEmacs CHISE[3]. When a user searches for the keyword “仏塔” on the site, the site finds the character “佛” relating to “仏” from the ontology data and searches both “仏塔” and “佛塔”. It partially includes AJAX (Asynchronous JavaScript + XML) technology to provide more convenient service.[4]

---

2) These products will be presented on the poster session.

3) XEmacs CHISE is a clone of XEmacs editor in order to represent every characters under a technology of “Ontology Engineering”. It is developed in Institute for Research in Humanities, Kyoto University.

4. The future

In the immediate future, we will try to modify the database based on the results of various Buddhist studies published after the publication of the Taisho-Daizokyo via our Web-based collaboration system. In order to manage this task, we will consider a means of version control for the fragments of the text to allow users to decide whether to adopt the latest or the former.

Bibliography

MORIOKA, Tomohiko. 2006. "Character processing based on character ontology", *IPSJ SIG Technical Report*, 2006-CH-072, pp. 25-32.

Nagasaki, Kiyonori, Takayasu Suzuki and Masahiro Shimoda. 2007. "The Development of a Collaboration System for the Text Database of the Taisho Tripitaka", *IPSJ SIG Technical Report, 2006-CH-70*, pp. 33-40.

Nagasaki, Kiyonori. 2007. "Digital Archives of Indian Buddhist Philosophy Based on the Relationship between Content Objects", *IPSJ SIG Technical Report, 2007-CH-75*, pp. 31-38.

Renear, Allen H. 2004. "Text Encoding", *A Companion to Digital Humanities*, Blackwell Publishing, pp. 218-239.

---

4) Tentative URL: http://21dzk.l.u-tokyo.ac.jp/SAT/satsearch.php

Abstract

The SAT Project and Web Solution

Nagasaki Kiyonori

International Institute for Digital Humanities, Japan

The SAT project, directed at present by Masahiro SHIMODA, was originated by late professor Yasunori Ejima in 1986 with the aim of building the text database of the Taishō Shinshū Daizōkyō, prior to the initiation of a similar project by the CBETA. These texts in this database, including around 6 million lines in 85 volumes composed of Buddhist canons compiled in India, China and Japan, have been checked character by character by Buddhist researchers by comparing them with the original texts of the first edition of the Taishō Daizōkyō, and with the database texts released by CBETA in 2005.

At first, our database was encoded in the Shift JIS system. At that time, missing characters were handled through the usage of the MOJIKYO-Font. Characters not contained in the MOJIKYO-Font were represented by the numbers of an original character code set. The structure of the database was based on the layout of the original Taishō volumes, using the structure of volume/page/paragraph/line. The first provisional version of the database, with some of texts left unchecked, was released in 2005 via the Internet.

At this point, we began to develop a viewer software program for our database based on the above format with the primary aim of facilitating the work of collaborating scholars. This program, able to represent the text in a vertical writing system and conducting rapid keyword searches, is implemented using Shockwave Flash, as this makes the system accessible to

users of various operating systems, such as Windows, Mac OS X, Linux, and so on.

Since early 2006, we have introduced a Web-based collaboration system on GNU/Linux in order to improve the efficiency of our work. This new system has enabled a number of Japanese scholars in separate geographical locations to engage in real-time collaboration. With the introduction of this system, we have shifted from the MOJIKYO-Font (which includes too many licensing restrictions) to "GT-Font", which, having been developed by academic bodies, is distributed with no charge for academic use. As for the characters not included in the Shift JIS and the GT-Font, — approximately ten thousand Chinese characters — we have created them in GT-Font style and are distributing them in the Web-based character database, which functions in complete harmony with the Web-based collaboration system.

In July 2007, we completed the task of correcting the wrong characters of the database and released the software to the contributors, with all the Indian characters, around ten thousand in all, installed. Now, we have started to work on releasing our text database on our Web site. We will publish it in XML format in October 2007. Our present policy is to publish the database in a format close to the original bound volumes. In order to follow open standards, we will change the format of our database. In addition, we are preparing a functional Web site on which users can search comparable keyword or display some fragments which are needed by users via formatted URI.

In the immediate future, we will try to modify the database based on the results of various Buddhist studies published after the publication of the Taishō Shinshū Daizōkyō via our Web-based collaboration system. In order to deal with this task, we will consider a means of version control to allow users to decide whether to view the latest text or the previous one.

논문투고일: 2010.11.30. 심사완료일: 2010.12.16. 게재확정일: 2010.12.18.