Journal of Digital Information, Vol 3, No 2 (2002)

Chinese Buddhist texts for the new Millenium - The Chinese Buddhist Electronic Text Association (CBETA) and its Digital Tripitaka

Christian Wittern
Institute for Research in Humanities, Kyoto University, Japan
Email: wittern@kanji.zinbun.kyoto-u.ac.jp

Abstract

This paper deals mostly with work by the Chinese Buddhist Electronic Text Association (CBETA) in preparing an electronic edition of a large collection of Chinese Buddhist texts. A short opening section about the history and significance of these texts is followed by a sketch of the digitization efforts prior to the CBETA project. The main part of the paper provides some background on the organizational structure and aims of CBETA, and then describes key aspects of the CBETA project, which are, among others: focus on quality assurance of existing electronic data, not foremost on input; reliance on open standards like the TEI Guidelines, XML and Unicode. The paper closes with an assessment of open questions, including the different formats for distribution of texts currently used by CBTEA, the reasons for their choice and the problems encountered. This will also touch on some more general questions concerning the distribution and continuing development of electronic ressources.

1 Introduction

Electronic resources have developed rapidly at many places during the last few years. Many problems, including the data format and structure, content delivery models, applications and interfaces have been encountered and solved.

Buddhist texts pose a somewhat unique problem on top of all this: since Buddhism came to span so many cultural and linguistic barriers over such a long period of time, the diversity of Buddhist primary sources is extraordinary. Precisely because of this, they also provide an ideal testbed for a digital library - any system that is able to deal with these complex and daunting issues will have little problem dealing with the other things out there.

The project described in this paper deals 'only' with one set of scriptures in this collection of canons, the Chinese Buddhist Canon that is used throughout East Asia. Although the main language is Chinese, there are citations, fragments and notes in other languages and scripts, including Sanskrit and Pali (in Romanized, devanāgari and siddham script), Tibetan, Japanese and others.

2 Chinese Buddhist Canon

Buddhist scriptures have been introduced to China and translated into Chinese over a period of more than 1200 years. Translation activity started in the Later Han period (in the 2nd century of the Christian Era) and continued with various degrees of intensity to about the end of the Song dynasty and the beginning of the Yuan dynasty (13th century).

When this translation process started, the Chinese had no idea that Buddhism had already developed into a number of competing schools in India, with sometimes contradicting teachings and commandments. Towards the beginning of the translation activity, any text that seemed interesting had been translated, with many translations being redone several times over the centuries as terminology and understanding developed.

It was at about the end of the 5th century that the growing corpus of translations of Buddhist texts from Sanskrit and Prakrit into Chinese had reached sizable proportions. Around this time, there was a need to organize the received scriptures, scrutinize the contents and asses whether they were authentic translations or mere fabrications of Chinese origin. Although the actual collection of scriptures into some sort of canon did not occur right away, the fact that they were recorded in bibliographic catalogs contributed to gradually having the scriptures considered as an entity, with heavily guarded entrance gates. At this point the collected translations were called `All of the sutras' (*yiqiejing*). One example from Dunhuang has the date of 479. About half a millenium later, during the Song period, first usage of the modern term *dazangjing* can be found, which is today commonly used to refer to the Chinese Buddhist Canon.

To preserve authoritative copies of the scriptures and prevent corruption, printing technology was adapted as early as the beginning of the 7th century, when a monk in a small monastery of Northern China embarked on a project to carve the most important scriptures in stone. The project was carried on over several hundred years and today we have access to about 14000 stone slabs, with hundreds of texts that comprise a major portion of the Chinese Buddhist Canon.

This remained an isolated project, however, and there was no plan that detailed the contents and sources of the carvings. It was therefore only with the beginning of woodblock printing in the 10th century that complete copies of the Chinese Buddhist Canon first became available. Shortly after the establishment of the Song Dynasty (960), work began in the remote province of Sichuan on imperial orders, which resulted in more than 1000 separate texts or more than 5000 scrolls being carved, printed and distributed all over the country. The first set was completed in 983. Since then, more than 20 new printing sets have been produced in China, Korea and Japan, each slightly differing in content and arrangement, although new admissions to the canon had been tightly controlled ever since the Song dynasty.

The oldest complete extant edition is that of the Tripitaka Koreana, the Korean edition of the Chinese Buddhist Canon (it should be noted that the Buddhist source texts used in China, Korea, Japan and even Vietnam are identical; together they form the tradition that gave rise to the Chinese Buddhist Canon). The woodblocks for this edition had been carved in the middle of the 13th century, comprising 1521 texts in more than 6500 scrolls.

The edition now most widely used as a standard reference to the Chinese Buddhist Canon is the *Taishō Tripitaka* (Taishō shinshū daizōkyō), edited by Takakusu and Watanabe (Tokyo, 1924-1932). This has been revised, rearranged and edited according to modern philological and text-critical principles, and it contains 3053 works in 85 volumes. While it does not contain all Buddhist scriptures of importance (there is a substantial volume of commentaries, records and historical texts in the *Supplement to the Chinese Buddhist Canon* (Zokuzōkyō), which have not been included in the Taishō Tripitaka), it has served as the textual source for many of the digitization projects that attempted to digitize the Chinese Buddhist Canon.

3 Digitization Projects: An overview

The last years have seen various efforts towards a complete digitization of the Chinese Buddhist Canon. Professor Lewis R. Lancaster of the University of California was among the first to realize the potential of digital texts and the enormous need for exchange, cooperation and standardization in this field. In 1993, he assembled delegates from various Buddhist electronic projects in different languages and scripts, and founded the Electronic Buddhist Text Initiative (EBTI) as a forum for exchange of information and sharing of technology among these projects. Subsequent meetings of the EBTI have been held at Haein-sa, Korea in 1994, Fokuang Shan, Taipei in 1996 and Otani University, Kyoto, Japan in 1997 and, together with PNC, ECAI and SEER at Academia Sinica, Taipei in January of 1999; a similar joint conference was held the following year at the University of California Berkeley in January 2000, while the conference in 2001 was a EBTI only meeting hosted by Dongguk University in Seoul, Korea.

There are of course plenty of other efforts, both modest and bold, small and large-scale in all countries that share the heritage of the Chinese Buddhist scriptures. The more important ones include different groups and individuals that the late Prof. Ejima Yasunori of Tōkyō University in Japan assembled over the years to help him realize his vision of an electronic databases of the Taishō Tripitaka. This project became a working group under the JAIBS (Japanese Association of Indian and Buddhist Studies) and was at times closely cooperating with Daizō Shuppansha, the publisher of the Taishō Tripitaka. The outcome of this cooperation was planned to be a series of CD-ROMs of the Taishō Tripitaka, one CD-ROM for each volume, despite the fact that technically the whole Tripitaka would easily fit on one single CD-ROM. This and some other factors that included high pricing and a failure to understand the special needs and features of the electronic medium led this venture to fail, and eventually it was discontinued after the publication of only four volumes. The group around Prof. Ejima has recently reassembled under the name SAT (see below).

At the small International Institute for Zen-Buddhism in Hanazono University, Kyōto, in the late 1980s Urs App conceived the Zen-Knowledgebase project. His aim was to put all information relevant to Zen-Buddhism, that is the original scriptures, commentaries, translations, bibliographies, maps, photos, video films and much more, into a connected computer database. This ambitious project began in 1990 with a 10 year plan. The present author joined that project in 1992 and contributed to its first widely known product, the 'ZenBase CD1', published in 1995. This CD-ROM could be seen as a snapshot of the project's workbench, with Zen-Buddhist texts released on it together with a number of research tools and utilities, designed to enable the researcher to make best use of these materials. Its main purpose had been to explore new and imaginative approaches to the electronic medium and to encourage others to do likewise.

Another important and bold step towards a complete version of the Chinese Buddhist Tripitaka was taken in Korea. Since Korea has the only extant complete set of woodblock, designated a Unesco World Heritage in 1996, plans were conceived to preserve these woodblocks by putting them into digital form, thus continuing the spirit of innovative technologies that surrounded the age of the carving in the 13th century, when the Korean peninsula witnessed printing with movable letters 200 years before this technique was applied by Johannes Gutenberg in 15th century Germany. The plan received encouragement and increasing support after the EBTI meeting in 1994 and with additional funding from a large company in Korea a first CD-ROM was produced by January of 1996. This CD-ROM tried faithfully to reproduce the image of the printed page, including vertical rendering; even the interlinear notes in smaller print were reproduced in the same way. Still more amazing, the great variety of character shapes that are found in woodblock prints are ported to the electronic text: over 30,000 different character shapes have been used in the roughly 50 million characters of this CD-ROM, compared to only about 12,000 in the modern Taishō edition, which has more than double the amount of text.

Taiwan has also seen a great number of efforts to help digitize the Buddhist scriptures. Due to technical difficulties and lack of coordination, there are many Buddhist electronic texts available on computer networks, but they vary in quality and are of rather limited scope. A considerable number of texts have been input by the Academia Sinica, but most of are available only to members. Some remarkable research tools have become available, however. For example, the excellent Fokuang Buddhist Dictionary has been put on CD-ROM in a very useful way and has been made available for a reasonable price. More recently, the collected works of Venerable Yin-Shun have also been made available for free on CD-ROM with a powerful search engine.

There have been some remarkable developments towards the creation of a complete full-text database of the whole Buddhist Tripitaka in Chinese. In January 1998, the above-mentioned group SAT, made up of Japanese Universities and members of the JAIBS, signed a contract with Daizō Shuppansha and was granted the right to create an electronic database of the Taishō Tripitaka and distribute it over the Internet. The project is scheduled to complete its task by 2006.

As a completely separate development, it became known in early 1999 that a Buddhist group in Hong Kong managed to put the whole Chinese Buddhist Tripitaka on a CD-ROM. From the table of contents it may be concluded that this CD-ROM contains material almost equivalent to all 85 volumes of the Taishō Tripitaka. Unfortunately, the text is not readable with software other than the reader provided, which requires a Traditional Chinese version of MS Windows. As in the case of the Korean Tripitaka, this software goes to great lengths to imitate the appearance of a book on the screen, but completely misses the new possibilities the electronic medium offers.

4 Realizing the CBETA Project

In February 1998, the Chinese Buddhist Electronic Texts Association (CBETA) was founded by Venerable Heng-ching, Taiwan University, and Venerable Hui-min, National Institute of the Arts, to coordinate efforts in Taiwan and promote the creation of a new scholarly digital edition of the Chinese Buddhist scriptures. The present author attended the founding meeting and joined CBETA in April 1998, working as an adviser to this project. CBETA was not going to start again with the input of Buddhist texts, but rather aimed at collecting and proofreading materials that had been put into electronic form elsewhere, thus ensuring high reliability throughout the database.

CBETA has received a grant from the Yin-Shun Foundation of North America and the initial plan was to release the complete canon of Chinese Buddhist scriptures (again according to the Taishō collection) consecutively within three to five years. Six volumes of the Taishō Tripitaka were released both on CD-ROM and on the Internet in December 1998. Today, 56 volumes of the Taishō collection are freely and publicly available through the CBETA Web home page. With about one year of the original work plan left, CBETA is now busy proofreading and integrating text-critical notes and other supplementary material that appeared in the original Taishō Tripitaka.

CBETA is making efforts to encode and markup the text using internationally accepted and widely used open standards like XML and TEI. CBETA is also cooperating closely with SAT on important issues like the representation of rare characters in the texts.

4.1 Origins

The CBETA project grew largely from a volunteer effort of people interested in the digitization of Buddhist texts. On a bulletin board forum hosted at Ven. Heng-ching's office at Taiwan University, discussions had been going on and even some preliminary tests completed, culminating in the release of the texts of Vol. 9 of the Taishō Tripitaka.

Inspired by the availability of the ZenBase CD1, which provided an example of how to achieve high quality digital texts even with a small team, it was realized that although data were becoming rapidly available, there was the need to gather these data and apply a quality assurance process.

To minimize the manpower and effort, it was decided from the start to rely heavily on supporting programs developed in-house and customized to the needs of team members.

4.2 Organizational structure

The organizational structure strongly reflects the origins of CBETA. Although the team was from the beginning divided into a number of separate groups, most groups started with only one or two members. As funding became available, more group members were hired; in the early years, most were added to the proofreading group, which worked on CBETA's core task. At the moment, there are the following groups:

- Research & Development
- Information Services
- Rare Characters
- Input
- Proofreading
- Network
- Distribution
- Bookkeeping

Of these, the Research & Development and Proofreading groups could be considered the core groups with most of the members of CBETA.

4.3 Methodology

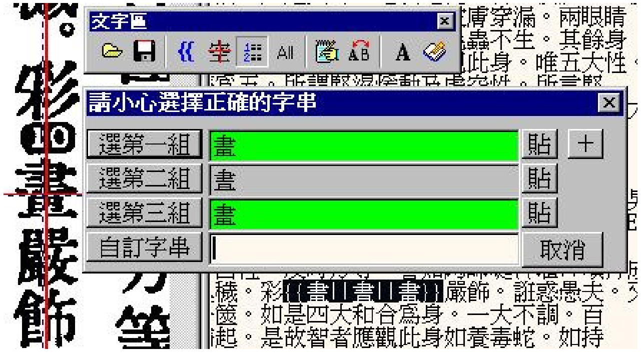
There are three areas in which the methods applied by the CBETA team might serve as an example to other similar projects:

- Computer assisted proofreading
- Consequent application of structural markup
- Systematic handling of non-system characters

Each of these items is discussed in more detail below.

4.3.1 Proofreading

As indicated above, CBETA tries to use information technology to minimize effort, manpower and cost while at the same time maintaining the highest quality standards. In the proofreading process this was achieved through a workflow that would try to optimize the time taken to find and correct errors in the sources. The proofreading process assumes that at least two, preferably three, electronic versions of a text are available. If this is not the case, the Input group is asked to prepare such copies as necessary. The electronic source files are then compared with a highly configurable program written in-house for that purpose and differences in these files are marked, separately for all input files. The proofreader than opens a program that displays a scanned image of the original and the electronic text side-by-side. Using this program, the proofreader can jump to the locations that have been marked and display the original page accordingly, as in an OCR proofreading system. This optimizes the most time-consuming process during proofreading, that is locating errors and finding out which version is correct.



選第一組：畫
選第二組：晝
選第三組：畫

穢。彩{{畫||晝||畫}}嚴飾。誑惑愚夫。
如是四大和合爲身。一大不調。百
是故智者應觀此身如養毒蛇。如持

Figure 1. Proofreading

The screenshot in Figure 1 shows a part of the interface for the proofreader. To the left is the scanned image, with the red haircross pointing to the character in question. The right part shows the result of the comparison of three files, with the one set of differences highlighted in black. Since some characters are difficult to discriminate on the screen, identical occurrences in two of the source texts are highlighted in the action dialog. It thus takes just one keystroke for the operator to make the necessary change in the file and move on.

There are obvious problems with this approach, for example it requires three independently created copies to be available, and the copies should also be created with different input techniques - it would not be useful to simply scan the same text multiple times and then use an OCR program to recognize the text. Identical errors in all three copies would not be identified. Preliminary tests with subsequent manual proofreading of the texts have shown that the results are as reliable as expected, with an error rate of approximately 1 in 10,000.

4.3.2 Markup

With ever more digital resources becoming available, we need to ask how can we expect these to be usable and useful after 10 years, let alone 100 or more years? There can be no guarantee, but one thing is obvious: we need to use open standards to shape our electronic texts. Texts that are available only through one specific software application will become useless in little time, since they are locked into systems that are becoming rapidly obsolete. The suppliers of digital resources need to conform to open standards, to provide texts and other material in a way that can be used with a multitude of software on a multitude of platforms, now and in the future.

This task also involves the interoperability of different electronic resources. In the same way that we can quote a book in any other book, we want to be able to connect different electronic resources, texts, dictionary entries, maps, biographies, historical or exegetic material, audio or video clips and so on. How can we achieve this? There is only one solution: again, all creators of electronic resources should use a common set of open standards.

What are open standards? Open standards are definitions of behavior for computer programs or of a format for computer data that are defined independently of any particular computer environment or vendor. Open standards have enabled the Internet to function with millions of computers interacting, where mainframes from the 1970s can exchange information with the latest supercomputers and notebook PCs without any problems.

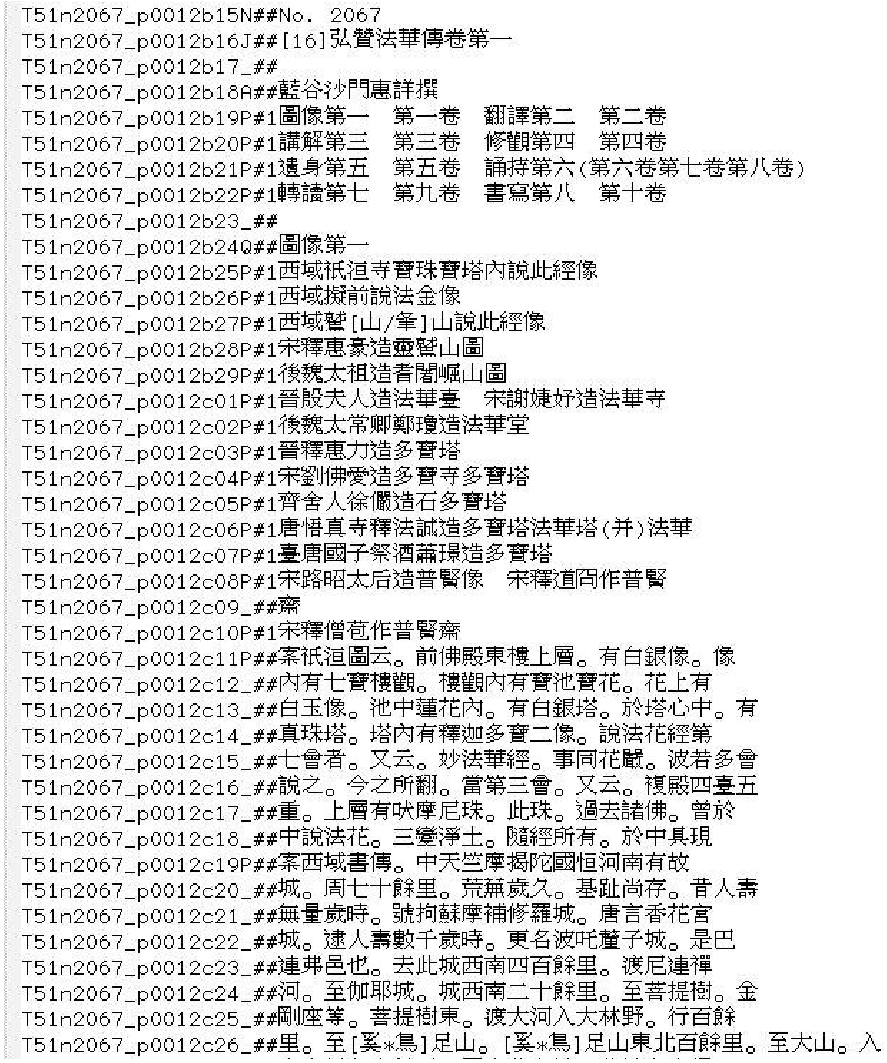
What standards do we need and who defines them? There is a standard designed for the markup of electronic texts and endorsed by the International Standards Organization (ISO) in 1986, defining the Standard Generalized Markup Language (SGML). This standard defines the syntax of markup and is used for example in the hypertext system that lays at the base of the graphical interface to the Internet, the World Wide Web.

The term `markup' originated in traditional printing, where markup are little marks or hints inserted in a text to tell the printer about headlines, font changes, and the like. With electronic texts, markup came to mean any way of inserting any kind of meta information into a text. This needs to be done in a systematic and standardized way, since it is intended to be used by computers, but it is also intended to be read and written by humans, so it cannot be too terse.

Using SGML as the syntax, an international group of more than 100 scholars from various fields of the Humanities formed the Text Encoding Initiative (TEI) and worked over more than seven years under the auspices of three learned societies to define some Guidelines for the Encoding of Electronic Texts [1], published in 1994. Today these guidelines are being implemented by a great variety of electronic text projects worldwide. There are of course specific needs in Buddhist texts that are not addressed here. EBTI is trying to give recommendations and develop guidelines for the specific needs of Buddhist texts and resources, but the work in this field is still in progress. Since these guidelines will only be useful if they address and solve all the problems found in encoding Buddhist texts, it is obvious that the more parties involved in the creation of such texts are involved in creating these guidelines, the better they will serve the purpose.

CBETA realized that using standardized markup for the creation of digital resources for Buddhist studies was a necessary condition for any further development of methodologies. Researchers will need to be able incrementally to add comments, definitions, pointers to related material as well as other meta information about a text. Markup, in combination with other knowledge representation strategies, can express the inherent information and retrieve it in ways that enable surprising new discoveries.

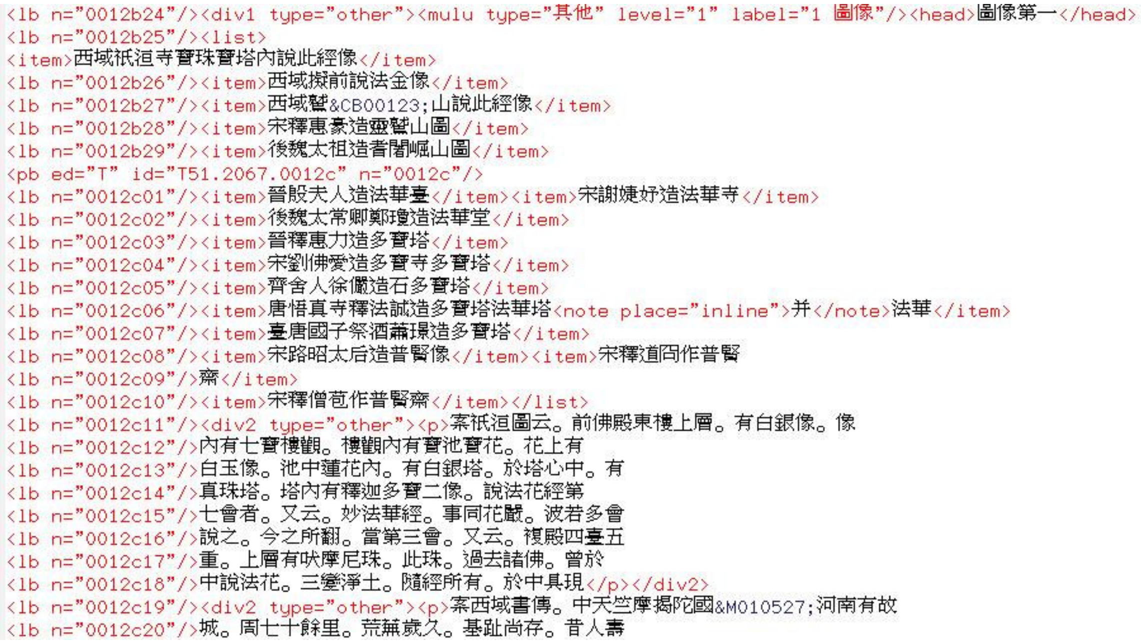
As I was the only one in the team with experience in the application of markup to electronic text when CBETA was established, I took it upon myself to convince the team members to use the TEI Guidelines as a base for this project. At the beginning, all markup was applied within the Research & Development group, but this turned out to be impractical. Since the proofreading group worked so closely with the texts, it was decided that in terms of efficient workflow it was the best place to apply markup to the texts. The tools in use by the group, however, would not allow use of SGML/XML and this would also place a heavy barrier to entry in terms of required skills (not to mention the language barrier, since at that time virtually no relevant documentation was available in English).



```
T51n2067_p0012b15N##No. 2067
T51n2067_p0012b16J##[16]弘贊法華傳卷第一
T51n2067_p0012b17_##
T51n2067_p0012b18A##藍谷沙門惠詳撰
T51n2067_p0012b19P#1圖像第一　第一卷　翻譯第二　第二卷
T51n2067_p0012b20P#1講解第三　第三卷　修觀第四　第四卷
T51n2067_p0012b21P#1遺身第五　第五卷　誦持第六(第六卷第七卷第八卷)
T51n2067_p0012b22P#1轉讀第七　第九卷　書寫第八　第十卷
T51n2067_p0012b23_##
T51n2067_p0012b24Q##圖像第一
T51n2067_p0012b25P#1西域祇洹寺寶珠寶塔內說此經像
T51n2067_p0012b26P#1西域擬前說法金像
T51n2067_p0012b27P#1西域鷲[山/夆]山說此經像
T51n2067_p0012b28P#1宋釋惠豪造靈鷲山圖
T51n2067_p0012b29P#1後魏太祖造耆闍崛山圖
T51n2067_p0012c01P#1晉殷夫人造法華臺　宋謝婕妤造法華寺
T51n2067_p0012c02P#1後魏太常卿鄭瓊造法華堂
T51n2067_p0012c03P#1晉釋惠力造多寶塔
T51n2067_p0012c04P#1宋劉佛愛造多寶寺多寶塔
T51n2067_p0012c05P#1齊舍人徐儼造石多寶塔
T51n2067_p0012c06P#1唐悟真寺釋法誠造多寶塔法華塔(并)法華
T51n2067_p0012c07P#1臺唐國子祭酒蕭璟造多寶塔
T51n2067_p0012c08P#1宋路昭太后造普賢像　宋釋道冏作普賢
T51n2067_p0012c09_##齋
T51n2067_p0012c10P#1宋釋僧苞作普賢齋
T51n2067_p0012c11P##案祇洹圖云。前佛殿東樓上層。有白銀像。像
T51n2067_p0012c12_##內有七寶樓觀。樓觀內有寶池寶花。花上有
T51n2067_p0012c13_##白玉像。池中蓮花內。有白銀塔。於塔心中。有
T51n2067_p0012c14_##真珠塔。塔內有釋迦多寶二像。說法花經第
T51n2067_p0012c15_##七會者。又云。妙法華經。事同花嚴。波若多會
T51n2067_p0012c16_##說之。今之所翻。當第三會。又云。複殿四臺五
T51n2067_p0012c17_##重。上層有吠摩尼珠。此珠。過去諸佛。曾於
T51n2067_p0012c18_##中說法花。三變淨土。隨經所有。於中具現
T51n2067_p0012c19P##案西域書傳。中天竺摩揭陀國恒河南有故
T51n2067_p0012c20_##城。周七十餘里。荒蕪歲久。基趾尚存。昔人壽
T51n2067_p0012c21_##無量歲時。號拘蘇摩補修羅城。唐言香花宮
T51n2067_p0012c22_##城。逮人壽數千歲時。更名波吒釐子城。是巴
T51n2067_p0012c23_##連弗邑也。去此城西南四百餘里。渡尼連禪
T51n2067_p0012c24_##河。至伽耶城。城西南二十餘里。至菩提樹。金
T51n2067_p0012c25_##剛座等。菩提樹東。渡大河入大林野。行百餘
T51n2067_p0012c26_##里。至[奚*鳥]足山。[奚*鳥]足山東北百餘里。至大山。入
```

Figure 2. Text file with `simple' markup

At that time, the standard format for the texts as used by the proofreading team was to have a location reference to volume, text number, page and line at the beginning of each line of the electronic text. We decided to extend this identifier by some columns and put shortcuts for structural markup there. It was easy to identify headings, footers, bylines, and so on. The paragraphs in our texts were clearly marked, so this information could simply be transformed into the `simple' markup[2]. Figure 2 shows the beginning of text number 2067 from volume 51. The last three columns before the Chinese text starts are used for the `simple' markup. Without going into detail, `P' will turn into markup for a paragraph, `Q' starts a new division, `A' is the author or translator, and so forth. If no markup is applicable, the `#' mark is used, the underscore is for lines that continue to belong to the previously mentioned markup entity. This *ad hoc* markup is then transformed to XML based on the TEI DTD with a perl program; all further editing is done on the XML files. The file generated from the above text is shown in Figure 3, with some additional editing applied.



```
<lb n="0012b24"/><div1 type="other"><mulu type="其他" level="1" label="1 圖像"/><head>圖像第一</head>
<lb n="0012b25"/><list>
<item>西域祇洹寺寶珠寶塔內說此經像</item>
<lb n="0012b26"/><item>西域擬前說法金像</item>
<lb n="0012b27"/><item>西域鷲&CB00123;山說此經像</item>
<lb n="0012b28"/><item>宋釋惠豪造靈鷲山圖</item>
<lb n="0012b29"/><item>後魏太祖造耆闍崛山圖</item>
<pb ed="T" id="T51.2067.0012c" n="0012c"/>
<lb n="0012c01"/><item>晉殷夫人造法華臺</item><item>宋謝婕妤造法華寺</item>
<lb n="0012c02"/><item>後魏太常卿鄭瓊造法華堂</item>
<lb n="0012c03"/><item>晉釋惠力造多寶塔</item>
<lb n="0012c04"/><item>宋劉佛愛造多寶寺多寶塔</item>
<lb n="0012c05"/><item>齊舍人徐儼造石多寶塔</item>
<lb n="0012c06"/><item>唐悟真寺釋法誠造多寶塔法華塔<note place="inline">并</note>法華</item>
<lb n="0012c07"/><item>臺唐國子祭酒蕭璟造多寶塔</item>
<lb n="0012c08"/><item>宋路昭太后造普賢像</item><item>宋釋道冏作普賢
<lb n="0012c09"/>齋</item>
<lb n="0012c10"/><item>宋釋僧苞作普賢齋</item></list>
<lb n="0012c11"/><div2 type="other"><p>案祇洹圖云。前佛殿東樓上層。有白銀像。像
<lb n="0012c12"/>內有七寶樓觀。樓觀內有寶池寶花。花上有
<lb n="0012c13"/>白玉像。池中蓮花內。有白銀塔。於塔心中。有
<lb n="0012c14"/>真珠塔。塔內有釋迦多寶二像。說法花經第
<lb n="0012c15"/>七會者。又云。妙法華經。事同花嚴。波若多會
<lb n="0012c16"/>說之。今之所翻。當第三會。又云。複殿四臺五
<lb n="0012c17"/>重。上層有吠摩尼珠。此珠。過去諸佛。曾於
<lb n="0012c18"/>中說法花。三變淨土。隨經所有。於中具現</p></div2>
<lb n="0012c19"/><div2 type="other"><p>案西域書傳。中天竺摩揭陀國&M010527;河南有故
<lb n="0012c20"/>城。周七十餘里。荒蕪歲久。基趾尚存。昔人壽
```

Figure 3. Text from Figure 2 converted to XML

4.3.3 Rare characters

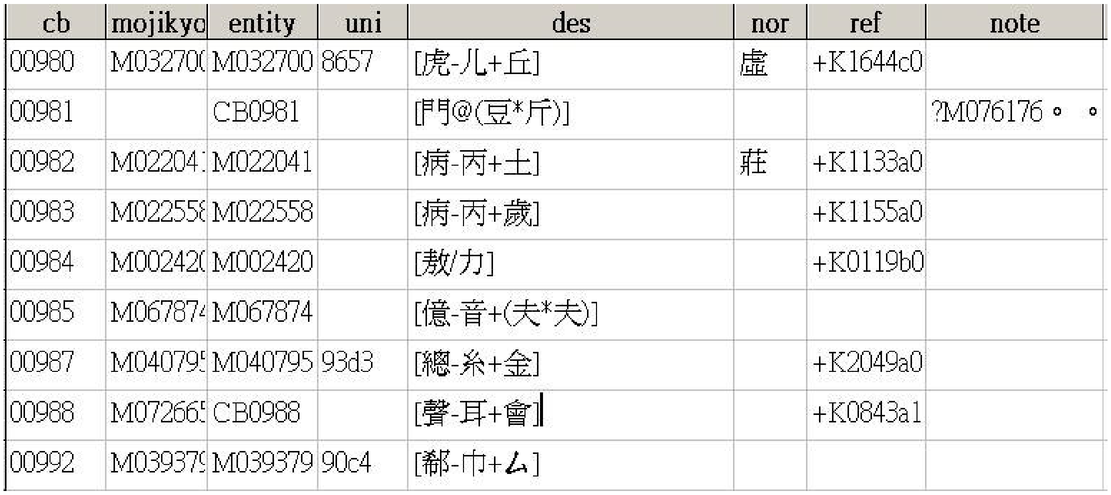
One of the biggest obstacles in the digitization of premodern East Asian texts in general and Buddhist scriptures specifically is the problem of premodern character forms.

I will try to illustrate the problem for those not familiar with the logographic scripts of East Asia. One of the features of logographic scripts is the fact that the smallest unit in the writing system, the character, is a semantic unit, but the orthography to express this unit depends on the time and place of the origin of a text. Modern character encodings for information processing are oriented towards the character forms in modern use and make it difficult to reflect the orthography found in the printed source forms in machine readable data. This is as if the encoding for English encoded words, not letters, thus making the coining of new words or the encoding of Middle English texts very difficult.

While the problem is significant in terms of the number of characters involved, the percentage of usage in a text is usually neglectible. More than 97% of any premodern text can be written with the most frequent ~5000 characters that are available in each of the national character sets in East Asia.[3] There are a number of well known approaches to this problem.[4] The baseline is that most characters can be expressed in the underlying standard character set (in the case of CBETA, this is the Big5 character set widely used in Taiwan and elsewhere for Traditional Chinese) and only a small percentage needs special treatment.

CBETA's approach to this problem is rather unique, in that it uses different methods in different phases of the workflow, depending on the tools available and the needs of the people working with them. In the first phase, during proofreading and inital markup, characters not found in Big5 are expressed using a simple algebraic language developed by CBETA. The advantage is that the character form can be guessed without requiring anything beyond plain text. At the same time, a list of such expressions was maintained. In this list, which has grown to more than 13,000 entries in the last four years, a serial number is assigned to each character and various other information is gradually entered; among other things, it was checked whether a modern character can be used as a replacement without distorting the meaning. During the conversion to XML, these glyph expressions were replaced with entity references, which is the usual way to handle characters that cannot be expressed in the document character set. This can be seen in Figure 2 where the expression [山/夆] on line b27 gets transformed to &CB00123 in Figure 3. For document delivery in various formats these entities can be expanded either with glyph-codes in custom fonts, image files for delivery over the Web, glyph expressions for plain text contexts and normalized forms for the general public and for search engines.

Figure 4 shows a few lines of the table CBETA is maintaining. The second column (numbers beginning with `M') is a reference to a large database of characters maintained in Japan; other columns have the corresponding Unicode codepoint (if available), the glyph expression, normalized forms and references to dictionaries. As can be seen, compatibility with Unicode is maintained by keeping this list of correspondences, so that at any time the whole database can be easily converted to Unicode.[5]



cb	mojikyo	entity	uni	des	nor	ref	note
00980	M03270	M032700	8657	[虎-儿+丘]	虛	+K1644c0	
00981		CB0981		[門@(豆*斤)]			?M076176 ∘ ∘
00982	M02204	M022041		[病-丙+土]	荘	+K1133a0	
00983	M02255	M022558		[病-丙+歲]		+K1155a0	
00984	M00242	M002420		[敖/力]		+K0119b0	
00985	M06787	M067874		[億-音+(夫*夫)]			
00987	M04079	M040795	93d3	[總-糸+金]		+K2049a0	
00988	M07266	CB0988		[聲-耳+會]		+K0843a1	
00992	M03937	M039379	90c4	[郗-巾+厶]			

Figure 4. Excerpt from CBETA's list of characters

5 Quo vadis - Electronic texts as source materials or consumer goods?

The aim of CBETA could be stated as 'to provide an electronic edition of Chinese Buddhist texts that is as accurate and reliable as possible, and that can serve as a foundation for further work both in Buddhist Studies and for Buddhist communities'. That is to say, the delivery not so much of a finished product, but of a raw material, a resource that can be used in the creation of a wide range of refined products.

The expectations of large parts of the intended audience were quite different, however. Almost nobody could make use of a naked XML file; angled brackets and entity references would obscure the text and would be considered noise by most readers. CBETA therefore had to develop a way to deal with these expectations. The result is that today the CBETA CD-ROMs (that are delivered free of charge, upon receipt of a letter including return postage) and the CBETA Web site contain the texts in a variety of different formats:

- Normalized plain text files (for reading in generic editors or file viewers)
- Normalized plain text files (with line endings adjusted so that they do not break words)
- 'Minimal size' plain text format (for use in handheld devices)
- HTML files
- HTML Help files (for Microsoft's documentation browsers; this includes moderate search functions)
- RTF version (for use in word processors, allowing easy printout of complete texts)
- XML version (for those who want to make use of it)

Additionally, PDF is currently under consideration. All these formats are generated automatically from the XML master files.

Although this sounds like a lot, these pre-formatted texts are still not fulfilling the needs of all users - for one thing, the display of non-system characters is hardwired and cannot be changed according to the user's wishes. More importantly, the underlying encoding is Big5 - some users in Japan or mainland China need different encodings. If all the different possibilities were pre-formatted, every single text of more than 2000 in this collection would have to be made available in more than 50 different formats.

This would be a maintenance nightmare and is clearly not possible. Another possibility that has been tested is to create the desired versions on-the-fly upon installation from the CD-ROM. While this allows the greatest flexibility and could deliver customized versions in four different encodings, the response from users was clear: nobody wanted to wait two hours, or sometimes up to eight hours on older machines, for the installation program to finish.

There is a larger issue that needs to be adressed. The current mode of distribution is based on a format derived from the XML files. Any annotations or edits a user might want to make to the files will be lost on the next update. There is also no easy way to allow users to enhance and enrich the text database with results of their research work.

CBETA is maintaining the source files in a CVS repository, and ways for users to access this repository directly are under consideration. This might ease the problem, but the infrastructure and expertise to make efficient use of this is not readily available for most users.

Maybe the biggest problem is that users expect a finished product that can be installed with a few mouse clicks, and which anticipates their needs before they are even aware of them. While this might be the prevailing trend in software applications and on the Web, we have to raise awareness that the availability of source texts in a format as versatile as XML make possible new ways of analysis and research, but no single polished and shiny graphical interface will support all needs.

References

- Chinese Buddhist Electronic Text Association http://www.cbeta.org
- SSaṃgaṇikīkṛtaṃ Taiśotripitakaṃ ('Digitization of the Taisho Tripitaka', shortened to SAT) is at http://www.l.u-tokyo.ac.jp/~sat/index.html.
- Tripitaka Koreana http://www.sutra.re.kr/english/default.asp
- IRIZ Web site http://www.iijnet.or.jp/iriz/
- TEI Consortium http://www.tei-c.org

Footnotes

[1] Sperberg-McQueen, C. Michael and Burnard, Lou (Eds.) (1994) *Guidelines for Electronic Text Encoding and Interchange*, Chicago and Oxford, sponsored by the Association for Computers and the Humanities (ACH), the Association for Computational Linguistics (ACL) and the Association of Literary and Linguistic Computing (ALLC).

[2] During the process of introducing this workflow, it was found that there was a need to fine-tune the markup beyond the unit of the line, since in some cases new paragraphs would not start a new line, so we later introduced some markers that were put inline into the text.

[3] A statistical breakdown of character usage of some Zen texts (about 3 million characters) can be found in Wittern, C. (1997) *Taming the masses. A practical approach to the encoding of variant and rare characters in premodern Chinese texts.*, presented at the Institute for Information Science of Academia Sinica in Taipei, Taiwan, March. The figures for the whole of the CBETA texts (80 million characters) are similar.

[4] See for example C. C. Hsieh, Chuang Derming and Lin Shih (1998) "A progress report of solving the missing character problem at Academia Sinica", presented at the 5th PNC meeting, May, Taipei, Taiwan, and published in *Proceedings of the Annual meeting of the Pacific Neighbourhood Consortium*, pp. 423-448; and Wittern, C. (1998) "The Issue of Rare Characters: Coding, Input and Output", presented at the 5th PNC meeting, May, Taipei, Taiwan, and published in *Proceedings of the Annual meeting of the Pacific Neighbourhood Consortium*, pp. 449-472.

[5] One of the most frequently asked questions is 'why don't you work in Unicode?' The answer has two parts: First, the suite of tools in the production line and the whole environment supports Big5 much better than Unicode. Second, Unicode is a moving target: recently the number of logographic characters has more than doubled and now exceeds 72,000. However, operating system and application level support is just starting and there are no reference works to identify characters within this enlarged Unicode. Conversion from earlier versions of Unicode would have been necessary anyway to keep up with the development, so using Big5 did not really make life harder.