Update on the Development of the Digital Dictionary of East Asian Buddhist Terms

Charles Muller, Toyo Gakuen University, Japan

[A full history and explanation of the content and structure of this dictionary was offered in a prior paper, entitled "The Structure and Function of the Interlinked Electronic CJK-English and Buddhist CJK-English Dictionaries" which is available on the Web at <http://www.human.toyogakuen-u.ac.jp/~acmuller/articles/dictionaries1.htm>. Therefore, aside from a brief summary, the present report will not repeat the explanations of structure and content contained in that paper, but focus instead on developments since the presentation of that paper —a period of approximately nine months.]

Summary

The compilation of the digital Dictionary of East Asian Buddhist Terms (DEABT), along with the companion Dictionary of East Asian Literary CJK Terms (DEALT) began in 1986. It was first converted to HTML format and placed on the web in 1995, at that time containing approximately 3000 terms. From 1995 to the present, the technical sophistication and structure of the dictionary have been under continuous development, with a gradual move toward XML validity. In the summer of 1998, both compilations were converted from S-JIS encoding to Unicode Text (UCS-2) format. In December of 1998, both compilations were made XML-valid, complete with rudimentary Document Type Definition (DTD) and eXtensible Style Language (XSL) files.

Present Web Version

On January 12, 1999, I uploaded a new edition of the online Dictionary of East Asian Buddhist Terms to my web site (http://www.human.toyogakuen-u.ac.jp/~acmuller). This was the first major re-publication of the dictionary in almost a year, the main highlight of which is the addition of approximately 700 new terms (going from 3500 to 4200). Some of the new terminology that is included has come from Iain Sinclair, who recently completed an M.A. on East Asian Esoteric Buddhism at the University of Western Sydney, and who has been sending materials derived from his research in that area. The bulk of the remaining new terms have come as a result of my translation work on Wonhyo's *Doctrine of the Two Hindrances* (*Ijang ui*), an East Asian Yogacara text.

The present publication of the dictionary is available as HTML files in two kinds of encoding: Shift-JIS and UTF-8. Since the dictionary is produced at the level of Unicode Text with MS-Word macros (rather than as HTML), and Word presently has no UTF-8 text editing function available, the source documents remain in UCS-2 and are subsequently converted to JIS and UTF-8 using the UNICONV conversion utility.

The DEABT has two new indexes: a full CJK index and a non-diacritical index of all terms. This means that you can search either index using the search tool on your browser. The purpose of the non-diacritical index is to allow users to type in search words from the keyboard, without having to worry about special input. For example, you can just type in "alayavijnana" or "drsti" and search.

The Move from HTML to XML

My largest headache during the past few years of presenting these dictionaries on the Web has been that of the conversion of the data source materials to HTML presentation format. Obviously, storing the material in HTML is not an option, as HTML is deficient in terms of necessary markup information. Database storage (such as MS-Access) has also proven impractical, as exportation and re-importation are always needed to perform global changes, and markup that could be done automatically in a word processor always needs to be hand-typed.

Therefore, I have decided to use XML/TEI as the primary storage/markup system. On the production of updated HTML versions of the dictionary, this material has been converted by macros into HTML. This regeneration of the materials usually takes a couple of hours —if the macros are fully and correctly set up before hand. But this is rarely the case, since during the several months duration between publication cycles, there are invariably numerous changes in the basic data structure, and new ideas for their presentation. This means that the macros invariably need to be rewritten each time, usually a two day task.

In view of these experiences, the long-awaited support of XML by the major browsers is a godsend to this project. For the first time, it is becoming possible to present the source data through a widely used browser application without having to run it through any sort of manipulative process. By simply constructing, and then making the necessary adjustments in a related style sheet file (an XSL file), it is now possible to present the materials directly from the source data file.

Therefore, upon the release of the Microsoft Internet Explorer 5 Beta 2 program in

December, I got down to the task of writing DTD and XSL files for each of the dictionaries, and then spent several painstaking days validating all of the data. Eventually I was able to make a rudimentary XSL-based presentation of the materials for IE5b2, which is available for public view at <http://www.human.toyogakuen-u.ac.jp/~acmuller/dicts>. As you can see if you look at these files, there is almost no style or useful functionality yet implemented. For example, no italics or colors, and no hyperlinking (XLinking). The reason for this is simply that (1) the MSIE5 beta is not yet fully supporting all XML functionality, and (2) the precise documentation on how to implement XML/XSL/XLink functions in the popular browsers is not yet available. Basically XML is still just too new (many of the final issues of XML implementation are still not resolved by the XML W3C working committee), but there is no doubt that rapid changes will be seen during the next six months or so, after which it will be feasible to present the dictionary in XML with at least the amount of functionality that was possible in HTML. This means that the present HTML versions of the dictionaries may well be the last of their type, as there will be little point in going to the trouble to continually regenerate these, when a single style file can serve the purpose.

Full Text Search, etc.

It would of course eventually be desirable to have the data presentable in a more dynamic manner, and to allow users more options for speedy search, as well as more direct usability of the dictionary in conjunction with their word processors and other research-related applications. I have actually made some serious attempts on my own at trying to add this kind of functionality, but have found the learning curve for the requisite programming skills to be just a bit much, as a scholar who desires to spend the bulk of his time studying and translating original Buddhist documents. Therefore, I have been concentrating my efforts on the content development of the dictionaries, and the creation of a clearly organized and subtly refined structure for the dictionaries using XML/TEI principles.

Content Contributions

At present, the only regular contributor of content to the DEABT is the above-mentioned Mr. Iain Sinclair (although there have been irregular contributions made from a number of scholars), but Mr. Sinclair' additions to the project in the area of Esoteric Buddhism have already added significant impetus to the project, and have helped in the development of the DTD. Mr. Sinclair was attracted by the dictionary during his use of it in the course of his thesis research. In the same vein, I hear, with continued greater frequency, from graduate students around the world who are using the DEABT for their thesis/dissertation research, and I suspect that there is a strong likelihood that future contributions will be coming from

these kinds of young scholars who are comfortable with working with digital resources, and who have had ample opportunity to acknowledge their value.

The dictionary is available for download through links on the index page. Contact and feedback from those who download is much appreciated.

Appendix:

Explanation of XML Markup for the Dictionary of East Asian Buddhist Terms (Updated 12/18/98)

XML/SGML/TEI Tags used inside the <sense> field of BDict. New tags may be added as necessary.

Note to Data Contributors: Since the tags external to <sense>, such as <entry ID>, and <resp> are generated automatically, you need not be concerned with these.

Automatically generated markup

<bdict> document file definition. File names are based on the traditional radical number
<strokes num=""> encloses the group of entries for characters that begin with the number of strokes after the radical
<chargroup char=""> encloses the group of entries that start with the same character

<entry ID="b001004E00-004084E58"> encloses a single dictionary entry. The id number is starts with either a "b" or "c", depending on whether it is the Buddhist or Classical Chinese dictionary. This is followed by a nine digit ID for each character in the term. The first three numbers indicate the traditional radical number. The second three indicate the number of strokes after the radical, and the final four are the Unicode hex number.

<head></head> Encloses the head word.

<pron></pron> Encloses the pronunciation group, which includes the following six areas:

<ch-py></ch-py> Chinese Pinyin
<ch-wg></ch-wg> Chinese Wade-Giles
<kr-hg></kr-hg> Korean Han'gul
<kr-mr></kr-mr> Korean McCune-Reischauer
<jp-kk></jp-kk> Japanese Katakana
<jp-rm></jp-rm> Japanese Romanization

It is expected that Vietnamese will eventually be included.

Content Area

All explanatory content for the head word.

<P></P>   Paragraph separator. Please use this to make paragraph breaks in individual entries, rather than word processor paragraph breaks.

A canonical collection, such as Taisho, Zokuzokyo, etc.

Emphatic. Used for boldfacing, italics, etc., in the case where no other indexing is appropriate.

Person's name

<person indlevel="1" lang="chn"></person>   Person's name, index level one, Chinese language. "Index level one" means that the tagged name is a direct equivalent to the headword in the language indicated. For Chinese, this would only be applied to the Pinyin reading.
<person indlevel="1" lang="ind"></person> Person's name, index level one, Indic language (Sanskrit or Pali not distinguished).
<person indlevel="1" lang="jpn"></person> Japanese language
<person indlevel="1" lang="kor"></person> Korean language
<person indlevel="2" lang="chn"></person>   Index level 2, which means that it is a significant term in the definition, but not analagous to the head word.
<person indlevel="2" lang="ind"></person> Same structure as above
<person indlevel="2" lang="kor"></person> Same structure as above
<person indlevel="2"></person>

<place indlevel="1" loc="chn"></place>   Place name, index level 1, Chinese location (and therefore Chinese language).
<place indlevel="1" loc="ind" lang="ind"></place> Indian location, Indic language. The language distinction is made here due to the fact that many Indian locations are also provided in Chinese translation and transcripion.

<place indlevel="1" loc="jpn"></place>   Japanese location (and therefore Japanese language)

<place indlevel="1" loc="kor"></place>   Korean location (and therefore Korean language)

<place indlevel="2"></place>   Place name, index level 2

<place indlevel="2" loc="ind"></place>   Indian location

<school indlevel="1" lang="eng"></school>The English language rendering of a school name.

<school indlevel="1" lang="ind"></school>

<school indlevel="1" lang="jpn"></school>

<school indlevel="1" lang="kor"></school>

<school indlevel="1" loc="ind" lang="ind"></school> An Indian originated school with its Indic name

<school indlevel="1" loc="ind" lang="chn"></school> An Indian originated school with its Chinese name

<school indlevel="2"></school>   School name, index level 2

<school indlevel="2" lang="chn"></school>

<school indlevel="2" lang="ind"></school>

<term></term>   Technical term

<term indlevel="1" lang="eng"></term>   Technical term, English language rendering

<term indlevel="1" lang="pali"></term>   Technical term, Pali rendering

<term indlevel="1" lang="skt"></term>   Technical term, Sanksrit rendering

<term indlevel="1" lang="tib"></term>   Techinical term, Tibetan rendering

<term indlevel="2" lang="eng"></term>   Technical term, index level 2

<term indlevel="2" lang="skt"></term>

<term indlevel="2" lang="skt"></term>

<text></text> Text name

<text indlevel="1" lang="chn"></text>   Text name index level 1, Chinese title

<text indlevel="1" lang="eng"></text>   English title

<text indlevel="1" loc="ind" lang="chn"></text> Indian text, Chinese title

<text indlevel="1" loc="ind" lang="ind"></text>   Indian text, Indic title

<text indlevel="1" loc="chn"></text>   Chinese provenance

<text indlevel="1" loc="jpn"></text>   Japanese provenance

<text indlevel="1" loc="kor"></text>   Korean provenance

<text indlevel="2"></text>   Text name, level 2

<text indlevel="2" lang="chn"></text>
<text indlevel="2" lang="eng"></text>
<text indlevel="2" loc="ind" lang="ind"></text>

<ref></ref>   Reference. This tag is applied to Chinese characters and compounds which do (or will) be listed in the dictionaries as a headword. These are converted to XML XLinks or HTML Hyperlinks as necessary.

<dictref></dictref>   Reference page number in other lexicon (Nakamura, Oda, Ui, etc.)
<school indlevel="1" lang="chn"></school>School (lineage) name, index level 1, Chinese language. Because of the multicultural nature of Buddhism, single schools are listed under various language names. Schools which are culture-specific (for instance, Popsong (Korean) and Nichiren (Japanese), should also have a Location identifier).

Entities (May be supplemented as necessary. In Unicode text format, those for which fonts are available are so encoded in the master source text)

&mdotblw;
&tdotblw;
&ddotblw;
&ndotabv;
&ndotblw;
&rdotblw;
&sacute; (Not needed for Unicode version)
&sdotblw;
&amacron; (Not needed for Unicode version)
&imacron; (Not needed for Unicode version)
&umacron; (Not needed for Unicode version)
&omacron; (Not needed for Unicode version)
&obrev; (Not needed for Unicode version)
&ubrev; (Not needed for Unicode version)