Moving into XML Functionality: The Combined Digital Dictionaries of Buddhism and East Asian Literary Terms

Project Manager's Report

Charles Muller
Toyo Gakuen University, Japan

I. Technical Review

The compilation of the Digital Dictionary of Buddhism (DDB) and Dictionary of East Asian Literary Terms (DEALT) began with my entry into graduate studies in Buddhist Studies, with the realization of the dearth of adequate lexicographical and other reference works in English language for the textual scholar of East Asian Buddhism in particular, and East Asian philosophy and religion, broadly speaking. I decided, during my first Buddhist and Confucian/Taoist texts readings courses to save everything I looked up, and have continued that practice down to the present, through scores of texts that I have studied. Although the content of these two lexicons is presently being supplemented by other interested parties, the terms that I have been compiling serve as the core of the work down to the present.

At the time that I began this process, I could not have dreamed of such a thing as the Internet, or even thought of the possibility of having this material available as a digital database--I was simply envisioning the eventual publication of a newer, larger, and more useful printed work. But as developments in the IT world progressed, the newly appearing potentialities gradually began to dawn on me. Then, in 1995, I tasted the Internet, and once I figured out how to insert <html> tags at the beginning and end of a text file, I was on my way to preparing these dictionaries for web publication--the first version of which I uploaded in the summer of 1995. It was not long after that that Christian Wittern discovered the DDB. He promptly downloaded all the files, and applied a basic SGML structure, which is the ancestor of the XML markup system used today.

Do to my lack of programming skills, as well as the limitations in the functions of the popular browsers, this framework mostly languished for a couple of years, during which time I would periodically regenerate new HTML versions of the dictionary with an array of Word macros. This was an extremely tiresome task, because my way of presenting the material was always changing slightly, and even slight changes usually necessitated a complete re-tooling of my macro system. Also, access to the data in the dictionaries was limited in method to hyperlinking, through an array of index files also generated from Word macros. The most important tool for making a dictionary really useful--a search engine, was throughout this time, absent. Thus, I longed for the day when I could just keep my data in some sort of stable, validated SGML/XML format, and have the data presented by users by a style sheet, or some other database-retrieval technique. Once XML support had been included in MS IE5 for a year or so, I went back and played with it some more, but still found that many aspects of XSLT were still not adequately supported, and with no Xlink/Xpointer support whatsoever, there was not much I could do, at least given my personal lack of the requisite programming skills to go beyond the level of support by popular browsers.

The first light at the end of the XML tunnel appeared in the summer of 2000, when Christian developed an experimental version of the DDB using the Zope system. This marked the first time that an attempt had been made to use the data in a form close to the original XML, and also a the first time a search engine had ever been applied to either of the dictionaries. As the maintainer of the dictionaries, however, this system presented difficulties to me in the sense that the data needed to be converted into thousands of small files--a situation that made the dictionary difficult for me to maintain locally. There were also problems with a lack of full native support for XSLT. Nonetheless, Christian's work marked the first time a version of the combined dictionaries had been generated more or less directly from the XML source, and was thus sufficient to make me decide to abandon the idea of the further headaches that invariably came with the generation of new HTML versions.

A major turning point in the history of the DEALT/DDB came this past January. While spending the New Year holiday in Kyoto I happened to be browsing a Japanese magazine on Palm computing, and noticed that Jim Breen's Japanese dictionary was becoming a sort of standard for inclusion on Japanese handhelds. It then occurred to me that although my data had always been publicly available, since the time of finalization and validation of the XML format, I had not really made an effort to let people in the IT world know that this data was freely available to download and program, the way Jim's data has been for a long time. So upon returning home, I immediately made an announcement of the availability of the data files on some of the major XML lists. Not long after, I was contacted by Michael Beddow.

II. The Most Recent Evolution: XML Comes Alive

Michael Beddow, a scholar of German Studies who has been involved in humanities computing since his undergraduate days in Cambridge in the mid 1960's, is a deeply experienced programmer, with a strong interest in using XML as a means of storage and delivery of literary and lexicographical documents. Seeing my announcement posted on Robin Cover's XML pages, he promptly downloaded the data and examined it. He contacted me shortly afterward to tell me that he was sure that he could program my data such that XSLT and Xlinking functionality could be produced in the latest versions of the standard browsers--and would I mind if he did that? (I didn't mind). Michael went to work, and within a week had, based on the markup structure of the DEALT, generated a complete array of indexes that used Xpointers to call up single-entry data units out of large files, each of which contained hundreds of entries. This was a landmark event for this project, because up to this time, if I wanted to call up a single-entry sized unit of data, the data files needed to be created to that size in advance. Or, using HTML anchors, I would link to a point in a larger file. With Michael's system, I could just plug my data into a system *as is*, and have it function like a real digital dictionary.

Of course, the whole process of development was not that simple. Michael knew about as much about East Asian languages as I did about Xpointers and Perl, so we spent about six weeks working with each other, rather intensively, to come up with a fairly workable system. My first reaction, after seeing the function of Michael's Perl-Xlinking system was to wonder (in an e-mail message) whether or not a search engine might not be applied using some of the same principles. As many of you know, devising a search engine that can deal with mixed Western/CJK text in UTF-8 encoding has been difficult to do so far, as the software has trouble parsing the divisions in the character codes. Within a few days, Michael also developed a prototype CJK-Utf-8 search engine.

Having accomplished this, there was one more function that I felt to be indispensable, in terms of the DDB, for taking best advantage of the value of the available data. While the present number of terms included in the DDB (8,000 at the time of writing) is certainly not small, it represents a very tiny fraction of the amount of terms, names, places, temples, schools, texts, etc., that are included in the entire Buddhist corpus. Thus, a search for a term conducted by someone whose research interests are significantly different than that of the compilers is likely to draw a blank screen.

Over the years, however, a group of scholars of East Asian Buddhism has been developing a comprehensive, composite index drawn from the indexes of dozens of major East Asian Buddhist reference works, which now includes almost 300,000 entries (described in further detail below). If the search engine could be made, after not finding an item in the DDB proper, to search this comprehensive index, the benefit to scholars would obviously be immense. Michael also added this function. Thus, in its present state, one may search for a term in the DDB, and on a "not found" condition, continue on to search this comprehensive index. In this case, the likelihood of not turning up one's search object is quite small (Michael has also written his own overview of the events described above, which you may peruse at *http://www.mbeddow.net/xml/technotes1.html*).

I have focused here on developments in the DDB, but please do note that all of the same technological enhancements have also been applied to the DEALT, except for the search through a comprehensive index, which, at present, has not yet been developed. We will give some concrete examples of web page format and search functions below, but first let us take note of some of the developments in the area of content.

III. Content Development

A. DDB

In January of 1999, the last time I presented at an EBTI meeting, I reported on a content update to 4,200 terms. At the writing of this present report, I am happy to inform you that that number has jumped to 8,000 and is continuing to increase rapidly. There are a number of reasons for this sharp increase in volume, the most fundamental being that of grant support from the Japanese Ministry of Education. We have, for the past three years, been receiving continuous support which has allowed for a number of things to take place which otherwise would not have. These are:

(1) Development of the Comprehensive Index -- During the two-year period of 1998-1999, we operated with a grant for the development of the comprehensive index (contents described in detail below, in the appendix). This project used the IRIZ Zendicts.dat file as a starting point (containing around 56,000 entries). To this we at Toyo Gakuen University, in collaboration with teams at Chung-Hwa and at IRIZ, added the indexes from a large number of major East Asian Buddhist reference works, bringing the total of entries up to almost 300,000. The use of this index in the task of developing the DDB is a tremendous advantage, as terms that are being looked up can

be located by computer instantly and certainly, right in the course of one's research. Thus, this index has been a great help in rapidly increasing the size of the DDB proper.

(2) Digitization of East Asian Reference Works -- The availability of the reference works themselves in digital format has also helped to increase our rate of expansion. Such lexicons as the *Fo Kuang Shan* dictionary and the *Ding Fubao* have already been formally and professionally digitized. We are adding to this by digitizing other valuable print works whose copyrights have expired, such as *Soothill's Dictionary of Chinese Buddhist Terms*, and works where we have permission to digitize from the copyright holder, such as Lancaster's *Descriptive Catalog of the Korean Buddhist Canon*. Students paid by our grants are scanning, OCRing, and correcting this data.

(3) Research Input from Graduate Student Assistants -- Also paid by a more recent grant, graduate students doing research in Japan are working for pay to input materials developed in the course of their research. While the volume of these materials has not been especially great, we have found this to be a good way to stimulate interest in the project. The students are also benefited by the chance to learn the computing techniques we are using for input, and to learn a bit about XML and so forth.

(4) Automated Input Technology -- Based on a set of indexes and tables, most of the assistants are able to use our system of MS-Word macros to add new entries rapidly. The macros create a ready-made entry structure, along with suggested readings of the entries for Chinese, Korean, and Japanese pronunciation. We are in the process of developing the necessary indexes to include Vietnamese as well. The shortcoming of this system is its limitation to MS-Word, but since the indexes upon which the system is based are all saved in Unicode text format, the development of an open platform input system which emulates our present Word system should be quite feasible.

(5) Input from Interested Scholars -- To date, we have received sizable glossaries from Iain Sinclair, Jamie Hubbard, Dan Lusthaus, and Gene Reeves. We hope that the continued increase in popularity of the DDB will stimulate the interest of more scholars to this end.

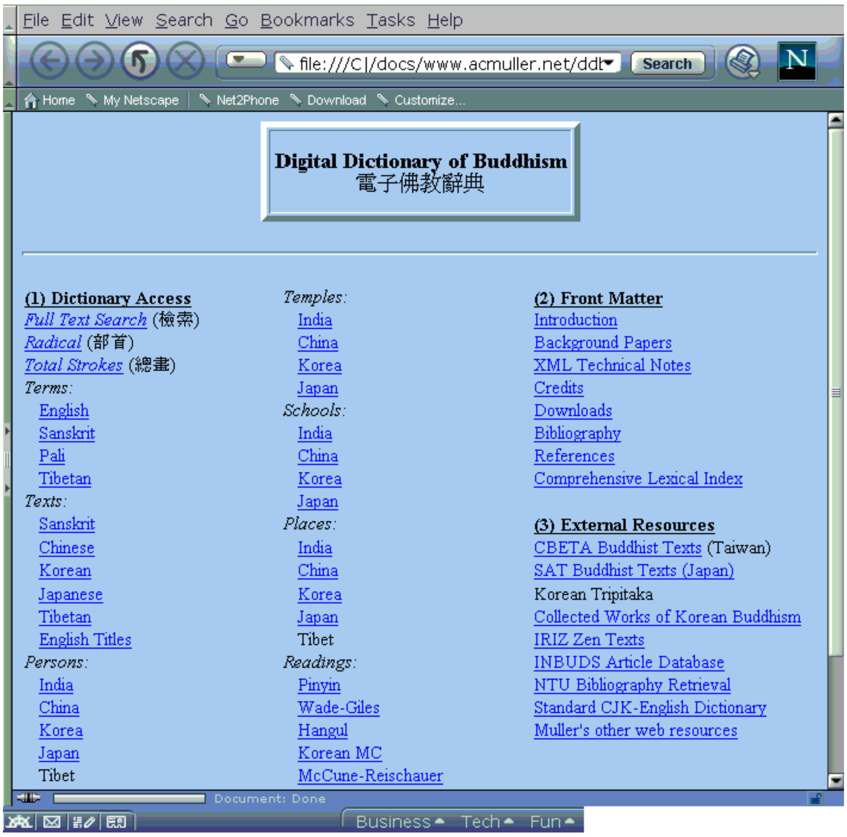
B. DEALT

The grants that I have applied for, along with the main focus of my efforts, have, for the past few years, been directed primarily at the development of the DDB, somewhat to the neglect of the DEALT. This one-sided emphasis however, is just a temporary state of affairs, and reflects nothing except a desire to bring one project to a certain level of completion before focusing on the next.

Nonetheless, in the course of my ongoing work with Buddhist texts, along with my frequent forays into Confucian and Daoist philosophy, I have been continuing to add compound words to the DEALT at a fairly rapid pace, and thus that collection now has almost 6,000 compound words. Also, fairly recently, I decided to make all 20,902 single character headwords in Unicode 2.0 available for browsing, even though only about 8,000 of these contain complete phonetic and semantic information. The rationale for this decision was that users who come across incomplete information might be influenced to do input. But I do not expect that the DEALT will ever get significant input in this way. At one point in the not-too-distant future, I will apply for a large grant for the specific purposes at making the DEALT as substantial a source of information as the DDB.

IV. The Present XML Browsing Environment of the Combined Dictionaries

In February of 2001, on the completion of Michael Beddow's draft setup of the combined dictionaries, I decided to open up my own domain name with a commercial web host, with the fear that the setup and maintenance of Michael's programming would be something that the small staff at Toyo Gakuen University could not handle. Thus the current home of the dictionaries is *www.acmuller.net*. This gateway offers a choice between entering the DDB and the DEALT. Upon entering the DDB table of contents page, the user is presented with the entire menu for the dictionary, including (1) the search engine and the various topic indexes; (2) the front matter and other explanatory materials for the dictionary, and (3) a small list of seminal resources for the study of classical East Asian Buddhist texts. As of Feb. 28 (just prior to the release announcement) this page was constructed as is shown below.


File Edit View Search Go Bookmarks Tasks Help
file:///C|/docs/www.acmuller.net/ddb
Search
Home My Netscape Net2Phone Download Customize...
Document: Done
Business Tech Fun

Digital Dictionary of Buddhism
電子佛教辭典

(1) Dictionary Access
Full Text Search (檢索)
Radical (部首)
Total Strokes (總畫)
Terms:
English
Sanskrit
Pali
Tibetan
Texts:
Sanskrit
Chinese
Korean
Japanese
Tibetan
English Titles
Persons:
India
China
Korea
Japan
Tibet

Temples:
India
China
Korea
Japan
Schools:
India
China
Korea
Japan
Places:
India
China
Korea
Japan
Tibet
Readings:
Pinyin
Wade-Giles
Hangul
Korean MC
McCune-Reischauer

(2) Front Matter
Introduction
Background Papers
XML Technical Notes
Credits
Downloads
Bibliography
References
Comprehensive Lexical Index

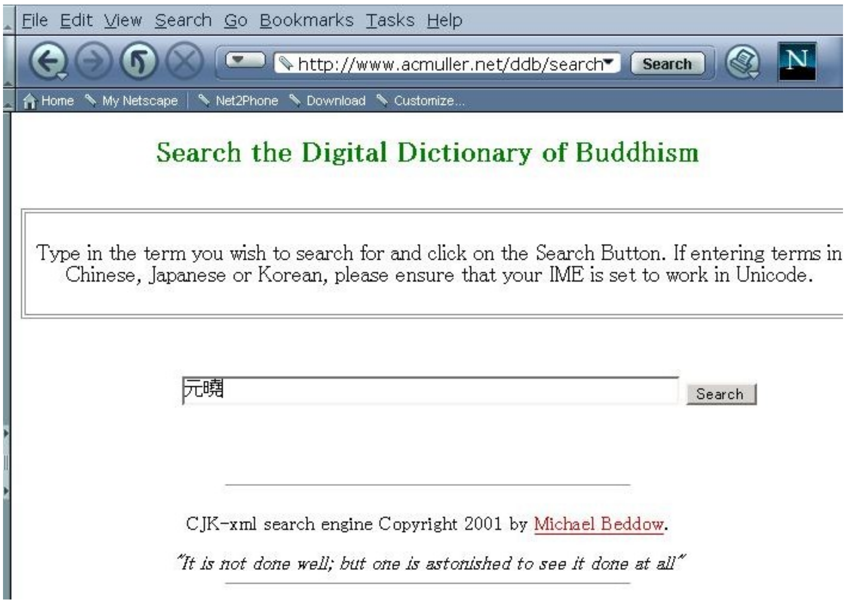
(3) External Resources
CBETA Buddhist Texts (Taiwan)
SAT Buddhist Texts (Japan)
Korean Tripitaka
Collected Works of Korean Buddhism
IRIZ Zen Texts
INBUDS Article Database
NTU Bibliography Retrieval
Standard CJK-English Dictionary
Muller's other web resources

figure #1: Table of Contents page -- DDB

As you can see, by presenting the entire dictionary menu, plus the most important scholarly sites for those doing research in East Asian canonical texts, this page becomes a useful one-stop portal for specialists in our area. Also, all links to areas within the site are done with absolute, rather than relative URLs. Thus, if you save this page to your desktop, you have ready access to all these materials any time you have an Internet connection.

We would imagine that most serious researchers and translators will use the search engine for basic access. But for those who are not sure of what they are looking, or who do not have a Unicode supporting IME, or who simply want to browse through the indexes as a form of study, the indexes still retain a good deal of usefulness.

At the time of the writing of this paper, the search engine still had some bugginess to it--most notably the behavior of turning up a white screen on the first try, but nonetheless functioning fine on the second try. The speed is not the fastest, but certainly not a long wait by any means. Since the web host is based in the U.S., American and European users should not have much trouble. At the time of writing, the search engine has the following interface:


File Edit View Search Go Bookmarks Tasks Help
http://www.acmuller.net/ddb/search
Search
Home My Netscape Net2Phone Download Customize...

Search the Digital Dictionary of Buddhism

Type in the term you wish to search for and click on the Search Button. If entering terms in Chinese, Japanese or Korean, please ensure that your IME is set to work in Unicode.

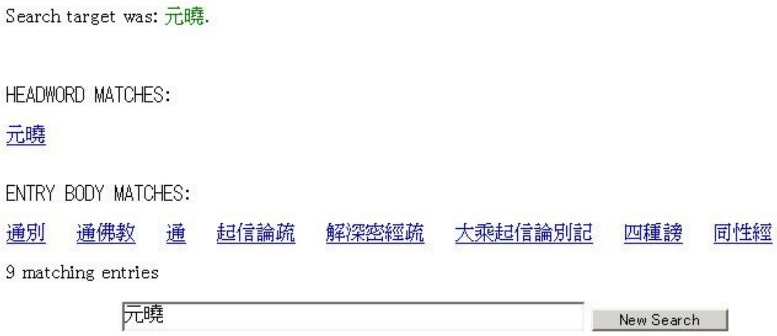
元曉 Search

CJK-xml search engine Copyright 2001 by Michael Beddow.

"It is not done well; but one is astonished to see it done at all"

Figure #2: Search Interface

When activated, the search will yield a menu of matches, containing headword hits, and instances occurring in the explanatory body of other entries, like so:



Search target was: 元曉.

HEADWORD MATCHES:

元曉

ENTRY BODY MATCHES:

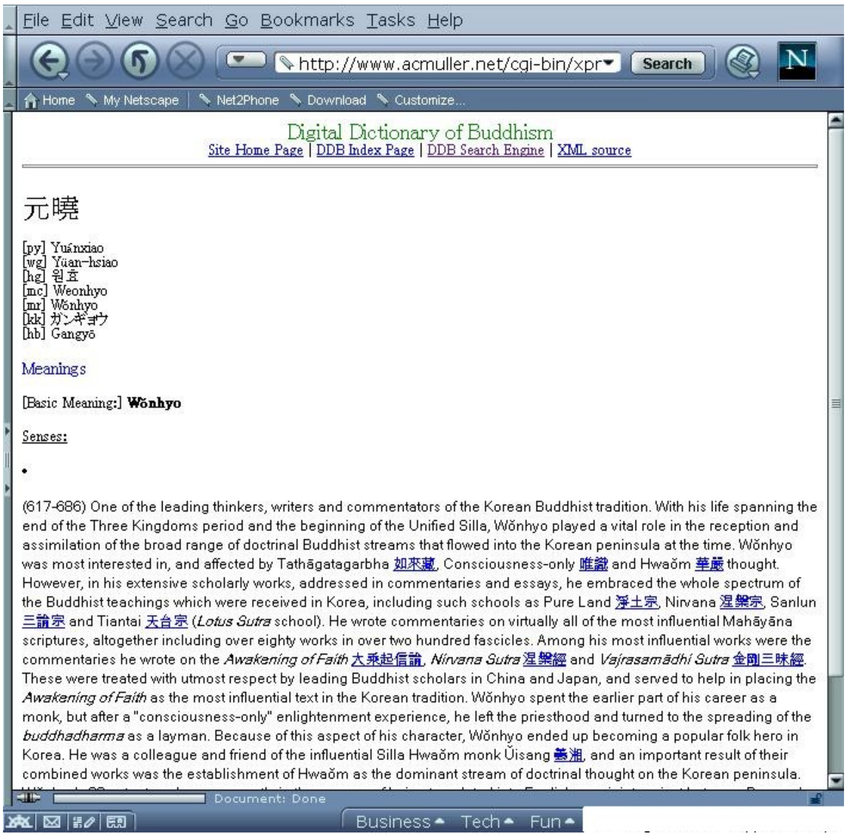
通別 通佛教 通 起信論疏 解深密經疏 大乘起信論別記 四種謗 同性經

9 matching entries

元曉 New Search

Figure #3: Headword and Text Matches

Finally, if one selects, for instance, the headword match, one can browse the term in question.


File Edit View Search Go Bookmarks Tasks Help
http://www.acmuller.net/cgi-bin/xpr
Search
Home
My Netscape
Net2Phone
Download
Customize...
Document: Done
Business
Tech
Fun

Digital Dictionary of Buddhism
Site Home Page | DDB Index Page | DDB Search Engine | XML source

元曉

[py] Yuánxiao
[wg] Yüan-hsiao
[hg] 원효
[mc] Weonhyo
[mr] Wŏnhyo
[kk] ガンギョウ
[hb] Gangyō

Meanings

[Basic Meaning:] **Wŏnhyo**

Senses:

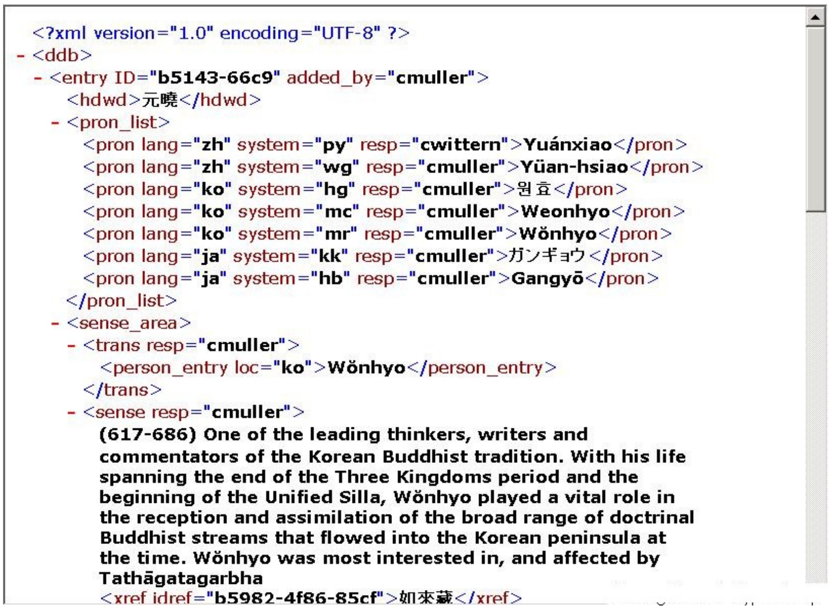
-

(617-686) One of the leading thinkers, writers and commentators of the Korean Buddhist tradition. With his life spanning the end of the Three Kingdoms period and the beginning of the Unified Silla, Wŏnhyo played a vital role in the reception and assimilation of the broad range of doctrinal Buddhist streams that flowed into the Korean peninsula at the time. Wŏnhyo was most interested in, and affected by Tathāgatagarbha 如來藏, Consciousness-only 唯識 and Hwaŏm 華嚴 thought. However, in his extensive scholarly works, addressed in commentaries and essays, he embraced the whole spectrum of the Buddhist teachings which were received in Korea, including such schools as Pure Land 淨土宗, Nirvana 涅槃宗, Sanlun 三論宗 and Tiantai 天台宗 (*Lotus Sutra* school). He wrote commentaries on virtually all of the most influential Mahāyāna scriptures, altogether including over eighty works in over two hundred fascicles. Among his most influential works were the commentaries he wrote on the *Awakening of Faith* 大乘起信論, *Nirvana Sutra* 涅槃經 and *Vajrasamādhi Sutra* 金剛三昧經. These were treated with utmost respect by leading Buddhist scholars in China and Japan, and served to help in placing the *Awakening of Faith* as the most influential text in the Korean tradition. Wŏnhyo spent the earlier part of his career as a monk, but after a "consciousness-only" enlightenment experience, he left the priesthood and turned to the spreading of the *buddhadharma* as a layman. Because of this aspect of his character, Wŏnhyo ended up becoming a popular folk hero in Korea. He was a colleague and friend of the influential Silla Hwaŏm monk Ŭisang 義湘, and an important result of their combined works was the establishment of Hwaŏm as the dominant stream of doctrinal thought on the Korean peninsula.

Figure #4: Headword Retrieval

For the user, this all looks and feels pretty much the same as it did in the earlier HTML versions of the DDB, but of course, what is happening is fundamentally different, as this HMTL text is being generated on the fly by Perl, XSLT, and Xlinking protocols.

The menu above provides standard returning links for returning to important places within the site, but also allows the user to view the XML source. This source view provides access to the names of those responsible for the various areas of the entry content--as below:



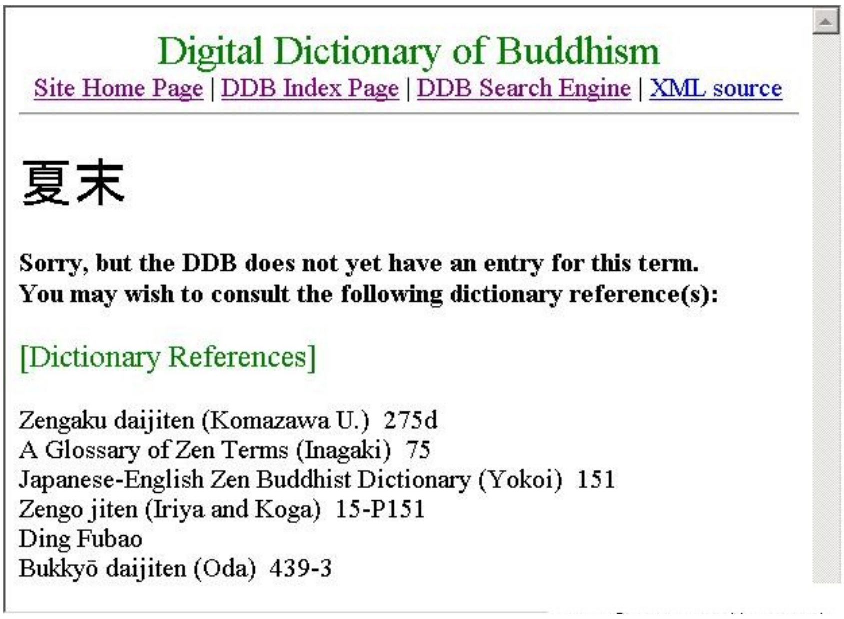
```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ddb>
  - <entry ID="b5143-66c9" added_by="cmuller">
      <hdwd>元曉</hdwd>
    - <pron_list>
        <pron lang="zh" system="py" resp="cwittern">Yuánxiao</pron>
        <pron lang="zh" system="wg" resp="cmuller">Yüan-hsiao</pron>
        <pron lang="ko" system="hg" resp="cmuller">원효</pron>
        <pron lang="ko" system="mc" resp="cmuller">Weonhyo</pron>
        <pron lang="ko" system="mr" resp="cmuller">Wŏnhyo</pron>
        <pron lang="ja" system="kk" resp="cmuller">ガンギョウ</pron>
        <pron lang="ja" system="hb" resp="cmuller">Gangyō</pron>
      </pron_list>
    - <sense_area>
      - <trans resp="cmuller">
          <person_entry loc="ko">Wŏnhyo</person_entry>
        </trans>
      - <sense resp="cmuller">
          (617-686) One of the leading thinkers, writers and
          commentators of the Korean Buddhist tradition. With his life
          spanning the end of the Three Kingdoms period and the
          beginning of the Unified Silla, Wŏnhyo played a vital role in
          the reception and assimilation of the broad range of doctrinal
          Buddhist streams that flowed into the Korean peninsula at
          the time. Wŏnhyo was most interested in, and affected by
          Tathāgatagarbha
          <xref idref="b5982-4f86-85cf">如來藏</xref>
```

Figure #5: XML Source Code Display

Those who have been watching the development of the DDB over time may notice the addition of a new field at the top of the <sense> area, called <trans>. This tag is borrowed from TEI, meaning "translation", but here referring strictly the word or short phrase as the direct common rendering that translators would use when rendering this term into English. The addition of this field was suggested by Christian Wittern, as a way of allowing the dictionary to operate as a translation tool, and to have a field that can used to quickly assimilate lengthy glossaries that give only short definitions. It has taken some work to go through the entire DDB and cull out the words that need to go in this field, but most have been done.

V. Inclusion of the Allindex Files

As mentioned above, one of the most important new developments of the DDB that hinges closely on the inclusion of the search engine, is the integration of the comprehensive composite index of East Asian Buddhological reference works. When a user's search does not find the term being sought, the search engine moves to search the *allindex* files, rendering up a listing of sources where one might find information on the previously searched term. For example at the time of the writing of this paper, the term *hamal* 夏末 ("end of the summer retreat"), was not yet contained within the DDB. But if we search for it, instead of drawing a blank, we will be given this information:



Digital Dictionary of Buddhism

Site Home Page | DDB Index Page | DDB Search Engine | XML source

夏末

Sorry, but the DDB does not yet have an entry for this term.
You may wish to consult the following dictionary reference(s):

[Dictionary References]

Zengaku daijiten (Komazawa U.)	275d
A Glossary of Zen Terms (Inagaki)	75
Japanese-English Zen Buddhist Dictionary (Yokoi)	151
Zengo jiten (Iriya and Koga)	15-P151
Ding Fubao	
Bukkyō daijiten (Oda)	439-3

There is more that needs to be said about the origins and trajectory of the Allindex project, but since it is somewhat of a separate topic, I have attached a discussion regarding it at the end of the paper as an appendix.

As can be seen, then, we are finally reaching a point where many of the impediments to full implementation are falling away. Most importantly, we are now starting to be able to handle Unicode encoded documents and take direct advantage of the power of XML. Of course, the function of the search engine, and the efficiency of the XSLT and Xlinking rendering methods will be continually improved as time passes. Our range of choice of Unicode-based font packages will also no doubt steadily increase. What is odd, however, is that in some ways, Microsoft is turning out to be somewhat of a hero in all this. For it is the Redmond giant which has thrown its resources into the widespread use of Unicode throughout its products, and which has, at least up to now, worked very hard to implement XML, and surprisingly, has been doing so according to the recommendations of the W3C. As these kinds of solid XML/Unicode based data sources proliferate, those whose software does not support it will find themselves increasingly at a disadvantage.

Thus, along these lines, one cannot search the DDB/DEALT without a Unicode supporting IME. And font packages that cover the whole range of international characters used in this compilation are presently only available for Windows systems. Hopefully, this situation will change in the near future. But in the meantime, how can we not but forge ahead and use the most advanced tools and support systems available to us?

Delivering CJK Dictionaries from Pure Xml Sources: A D eveloper's Perspective

Michael Beddow

Probably the most important thing to stress about the collaboration between Charles Muller and myself on an xml-based delivery platform for the DDB and the DEALT is that no more than six weeks elapsed between our first contact and the announcement of a fully-functional system (and indeed one that had more functions than either of us had envisaged at the start). Perhaps even more noteworthy is that I had the core of the system up and running (in so far as individual entries were being retrieved from the larger files) within a single day of first downloading the data.

I say this not to praise myself as a lightning-speed programmer, but to bring out what is it that makes xml such a hugely important force for changing the way we in the Humanities work with digital data. Of course, years of effort had gone into Charles Muller's collection and mark-up of the data, and months of work had gone into my development of the modules from which I built a delivery platform tailored to that data; but because data marked up in xml really does describe its own structure, and because software that follows the recommendations for processors issued by the W3C is intrinsically adaptable to any sort of well-formed xml, none of our earlier independent work had to be redone to get the data on-line. When recoding his data in xml, Charles had been focusing on the scholarly content and the abstract structure, with relatively few detailed ideas about how it would eventually be delivered to users (who in the meantime continued to access his work via the conventional html site). For my part, I had been working on techniques of retrieving fragments of larger xml documents and rendering them into html on demand, with no substantial experience either of handling CJK data or of the problems specific to lexicographical applications. Yet, once an announcement on xml-doc brought us together, the required retrieval, delivery and rendering system more or less sprang into life of its own accord. "Self-describing data" pretty much engendered a self-creating delivery system. It was an exciting, if slightly uncanny, experience.

The old and the new

One of the chief benefits Charles had foreseen when moving to xml encoding of his material was the eventual possibility of using XLink and XPointer[1] technologies to allow users to

[1] Of the many explanations of XLinks and XPointers available on line, the one that to my mind strikes the best balance between comprehensibility and depth of coverage is at http://www.javacommerce.com/tutorial/xml/linking.html

retrieve selected fragments of larger documents. In an html implementation, either the editors have to maintain a very large number of small documents, with all the version management problems that entails, or users have to accept that the results of their queries are large documents of which only a small portion may be relevant to what they were looking for. Anticipating the removal of this serious limitation implicit in html, Charles Muller had been marking up internal and external links in the xml version of his materials using a basic form of XLink/XPointer notation, but had believed that these links would only be used as intended once browser (and server) support for XLinking was widely implemented.

I was able to show that by using some simple cgi scripts in combination with server-side xslt transformations, it is possible to implement a small but useful subset of the (still not finalised) XPointer and XLink proposals that can be used with present day browsers and servers. I originally developed these techniques, based on freely-available Open Source models, to allow the on-line publication of a long monograph of mine from a single canonical and easily-maintained xml file, while enabling users to request and receive portions of this single file as small as a single (printed) page, transformed on demand from the TEI-conformant xml into html.[2]

Like the html-based system that preceded it, the xml-based platform involves the creation of many thousands of files, largely because of the caching and indexation facilities it uses to speed retrieval and delivery. But there is an immensely significant difference from the editorial point of view. The thousands of html files had to be maintained by the editors themselves; in my system, all the editors need concern themselves with are the core xml files into which they enter their data. There are many other supporting files; but they are invisible to both the end user and the resource authors, and are generated and maintained transparently by the underlying system. The authors create and maintain xml files of whatever size best suits their methods of working, and whose structure is determined by their scholarly analysis of the material. The system validates, partitions and indexes those files, allows users to locate the items within them that they need, and renders the retrieved items into web pages for delivery, creating hyperlinks for any internal or external cross-references as it does so.

About half way through my work on automatically creating the existing indices, Charles asked whether it would be feasible to create a freetext search engine that would supplement

The current W3C proposals (not easy reading) are at http://www.w3.org/TR/xlink/ and http://www.w3.org/TR/xptr

[2] This work may be seen at http://www.mbeddow.net/foh/

these indices as a means of access, and for some classes of user maybe even replace them. I replied that this would be an interesting challenge, but not one that could be met in the short term. Later that same day I was having lunch in a country pub, when I had an idea which, impolitely neglecting my companions, I scribbled on a beermat. That evening I translated the scribbles into program code and found I had a half-way usable search mechanism.

This free-text search engine is still to this day only half-way usable. Some of the problems lie in my own coding, which needs, and will in due course receive, much more work, others stem from aspects of the underlying system libraries which only come to light when complex regular expressions involving utf-8 encoded characters from across the entire Unicode range are let loose on multilingual texts. There is also an irritating bug, alluded to in Charles Muller's paper, which occurs only on the (FreeBSD) hosting server but cannot be reproduced on my (Linux) development system, and which causes the initial failure of some search attempts. I hope users will not find it distractingly flippant that, as a much-needed caveat-cum-apology, I have cited on the query form the remark I suspect the father of Anglophone lexicography might have made about my efforts, had he encountered them betwixt his observations of women preachers and dogs walking on their hind legs. Given Dr Johnson's place in scholarship, this seemed more appropriate than the other citation which also springs to mind in my defence, G.K Chesterton's observation that "if a thing's worth doing, it's worth doing badly".

Aside from the search engine, the other thing the new system brings from a user's perspective is a more commodious display of the data. The layout, ordering and indeed the contents of the delivered html can easily be changed by editing a single controlling xsl style sheet, without touching the xml data itself, so it is easy to act on user comments (or editorial second thoughts) about the presentation of the material which previously might have required the recreation of thousands of separate html pages. In other words, the separation of visual design from logical structure that xml allows for is here given full scope.

The nature of xml markup has also allowed a significant extension what the user specifically of the Digital Dictionary of Buddhism can be offered. Because the very large set of references to Buddhist CJK terms in printed or other digital dictionaries which Professor Muller and his associates have assembled were also marked up in xml, the DDB's facilities could be greatly expanded with very little programming effort. If a user looks up a term in the search engine which is not in the DDB (or if s/he follows a provisional cross-reference in the DDB where the reference target has not yet been edited into place), a secondary lookup is performed on the external references data. If the term concerned is found there, the user is offered a listing of the locations in those external sources where the term is defined or explained. Because of the very large number of entries in this secondary data collection (c.300,000 and rising) , lookups are assisted by a Berkeley db database (itself automatically built from the core xml) interposed between the client and the xml sources: this is the only instance in the current implementation where information is not located by a direct parse of the core xml files.

The system in operation

Each headword in the dictionary has a unique identifier (id) as part of the markup. This id is derived algorithmically from the name of the dictionary plus the Unicode numerical representation of the characters in the term. When a term is requested, either from one of the various user-accessible indices or as a result of a search engine query, the relevant id is passed to a cgi script on the server. That script parses the appropriate xml file[3], locates the entry by its id and extracts it, then passes the resulting xml fragment on to an XSLT processor[4], which converts it into html while building the necessary hyperlinks for any cross references the entry contains.

In practice, this process is complicated (but also accelerated from a user perspective) by a system of caching, by which both xml fragments and the corresponding html version, once created by an initial request, are stored so that future requests can be met without further parsing or transforming, until the editors alter the items concerned in the xml (which automatically invalidates any cached copies of the altered material), or alter the xslt style sheet that controls presentation (upon which all cached html is marked invalid so that it will be regenerated with the changed presentation next time the xml is retrieved).

Platform requirements

Though earlier work on xml fragment retrieval and rendering in real time was done on University servers which I specified and managed, giving me complete control of the hardware and software, the programs that deliver these dictionaries can run on servers which offer only the limited configuration facilities found at the inexpensive end of the commercial hosting market. No privileged access to the machine is needed to install or maintain them. There is, of course, a performance penalty: the whole thing would run faster and handle more simultaneous users without performance deterioration if it could be moved "in process" with the Web server, so that the script handling system did not have to be loaded and initialized for every single

[3] The parser is expat, originally by James Clark, now maintained by Clark Cooper and Fred L Drake, Jr, available from http://sourceforge.net/projects/expat/

[4] The processor used is Xalan C++ Version 1.1 from http://xml.apache.org/xalan

request, as happens at the moment. But performance is broadly satisfactory for the present size of the datasets and should cope with their planned expansion. And my modified methods mean that other scholars who would like to deploy a version of this system adapted to their particular data have the prospect of getting it to run without excessive dependence on the co-operation or expertise of their local server administrators. One indispensable requirement, for CJK applications at any rate, is the presence of up-to-date system libraries for handling Unicode, and experience suggests that these are more commonly found on commercial sector servers than on campus facilities.

The Moral of this Tale

Until mid January 2001, Charles Muller was to me just a name on a list of translation resources I maintain. And he, understandably, hadn't heard of me at all. But thanks to the Internet we were, within a matter of days, pooling our knowledge and interests and working together across a distance of nine time zones as effectively as if we had been in neighbouring offices, with results that I hope speak for themselves. Humanities scholars who still insist that computers are no more than glorified but temperamental typewriters and campus finance officers who believe only scientists need decent computer hardware or network connections might like to consider revising their views. And anyone who thinks XML is either just a fad or tomorrow's technology can see its enabling power at work right here and now.

Appendix: The Allindex Database

Composite Index of East Asian Buddhist Lexicographical Sources ("allindex.xml")

Primary Compilers: Urs App, Christian Wittern, Charles Muller, Michel Mohr, Hur In-Sub

Initial Release Date: 4/26/99

Updated: 03/01/2001

The "allindex" file is an ongoing compilation of the indexes of East Asian dictionaries of Buddhism. It was initiated in the form of the Zendics.dat file published on the *ZenBase CD-ROM*, by the International Research Institute for Zen Buddhism (IRIZ), developed by Urs App, Christian Wittern, and their staff. That file contained complete index information for the sources listed in the IRIZ bibliography below (58,563 entries). Using this file as a basis, we have been continuing to add indexes from other lexicons, among the most significant of which are the index to Nakamura Hajime's *Bukkyōgo daijiten*, the *Fo Kuang Shan Dictionary*, *Ding Fubao*, Hirakawa's *Buddhist Chinese-Sanskrit Dictionary*, the *Bussho kaisetsu daijiten*, the Oda and Mochizuki dictionaries, as well as many other smaller Buddhist dictionaries. This file will be further supplemented by the digitization of other East Asian Buddhist lexical resources, presently in progress.

Each entry contains a field for the headword (Chinese characters), the Chinese, Korean, and Japanese readings of these characters, an ID number, and a list of sources for the word that have been identified.

[Sample]

```
<entry ID="b4e00">
        <hdwd>一</hdwd>
        <pron lang="zh" system="py">yī</pron>
        <pron lang="ko" system="hg">일</pron>
        <pron lang="ja" system="hi">いち</pron>
        <dictref>
                <dict name="ZGD">28a</dict>
                <dict name="Ina-Z">75</dict>
                <dict name="ZD">269</dict>
                <dict name="Naka">45a</dict>
                <dict name="FKS">1111</dict>
                <dict name="BCS">0001</dict>
                <dict name="YBhI"/>
        </dictref>
</entry>
```

Indexed works

1. From IRIZ
  a. Urs APP and Christian WITTERN

Daitō shuppansha. 1979 (rev. edition 1994). *Japanese-English Buddhist Dictionary* 日英佛教辭典. Tokyo (page numbers and terms of both editions are included). [JE]

Genkyō Zenji 元恭禪師. 1908. *Zengaku zokugokai*. 禪學俗語解. Tokyo: Kaiunji 海雲寺 (Republished in 1991 by the Zenbunka kenkyūjo as part of the *Zengo jisho ruiju fu saku in*. 禪語辭書類聚 付索引. Kyoto: Zenbunka kenkyūjo 禪文化硏究所).

Inagaki, Hisao 稲垣久雄. 1991. *A Glossary of Zen Terms*. Kyoto: Nagata Bunshōdō 永田文昌堂. [Ina]

Iriya Yoshitaka's 入矢義高, trans. *Baso no goroku* 馬祖の語錄. Footnotes to Iriya's Japanese translation of the *Mazu yulu*: Kyoto: Zenbunka kenkyūjo 禪文化硏究所, 1984.
Iriya, Yoshitaka 入矢義高, and Koga, Hidehiko 古賀英彦. 1991. *Zengo jiten* 禪語辭典. Kyoto: Shibunkaku 思文閣. [ZGo]

Komazawa daigaku nai Zengaku daijiten hensansho 駒澤大學內禪學大辭典編纂所. 1977. *Zengaku daijiten* 禪學大辭典. Tokyo: Taishūkan shoten. [ZDG]

Miura, Isshū, and Fuller Sasaki, Ruth. 1966. *Zen Dust*. Kyoto: The First Zen Institute of America in Japan (Out of print).

Mujaku Dōchū 無著道忠. 1979. *Kattō gosen* 葛藤語箋. Kyoto: Chūbun shuppansha 中文出版社. (pp. 868-1100 of volume 9 下 of the *Zengaku sōsho* 禪學叢書 edited by Yanagida Seizan 柳田聖山). Our index also contains the page numbers of another edition: *Kattō gosen* 葛藤語箋. Tokyo: Komazawa University's Compiling Office of the Zen Dictionary, 1959.

Mujaku Dōchū 無著道忠. 1979. *Zenrin shōkisen* 禪林象器箋. Kyoto: Chūbun shuppansha 中文出版社. (Volume 9 上 of the *Zengaku sōsho* 禪學叢書, edited by Yanagida Seizan 柳田聖山). Our index also contains the page numbers of another edition: Zenrin shōkisen. Kyoto: Seishin shobō 誠信書房, 1963.

Nakamura, Hajime et al. 中村元など. 1989. *Iwanami Bukkyōjiten* 日英佛教辭典. Tokyo: Iwanami. [Iwa]

Shibayama Zenkei 柴山全慶. 1972. *Teihon zenrin kushū* 定本禪林句集. Kyoto: Kichūdō 其中堂.

Yanagida Seizan 柳田聖山 trans. *Rinzairoku* 臨濟錄. Footnotes to Yanagida's Japanese translation of the *Linji lu.* Tokyo: Daizō shuppansha, 1972.

Yokoi, Yūhō 横井雄峯. 1991. The *Japanese-English Zen Buddhist Dictionary* 日英禪語辭典. Tokyo: Sankibōō Buddhist Bookstore. [Yo]

Yuan Bin 袁賓. *Chanzong zhuzuo ciyu huishi.* 禪宗著作詞語釋. Shanghai: Jiangsu guji chubanshe 江蘇古籍出版社, 1990.

Zen no goroku 禪の語錄 . The footnotes to all 17 vols of the series published by Chikuma shobō: Tokyo.

b. At IRIZ; ABE Rie, Urs APP and Michel MOHR

Mochizuki Shinkō. *Bukkyō Daijiten.* [MZ]

Oda Tokuno 織田得能. *Bukkyō Daijiten* 佛教大辭典. Tokyo: Daizō shuppan kabushiki

kaisha, 1995. [Oda]

---

2. At Toyo Gakuen University; Charles Muller, et. al.

Nakamura Hajime 中村元, ed. *Bukkyōgo daijiten* 佛教語大辭典. Tokyo: Tokyo shoseki, 1985. -- {digitized by Charles Muller and Maki Miyaji} [Naka]

Dongguk University Research Center for Buddhist Culture, ed. *Kankoku bukkyō kaidai jiten* 韓國佛解題辭典. Tokyo: Kokusho kangyōkai, 1982. -- [KBKJ] {digitized by Charles Muller}

Ono Gemmyō 小野玄妙, ed. Bussho kaisetsu daijiten 佛書解說大辭典. Tokyo: Daitou shuppansha, 1999. {digitized by Charles Muller and Maki Miyaji} [bsk] [index=bski]

Saito Akitoshi 齋藤昭俊 and Naruse Yoshinori 成瀬良德, ed. *Nihon bukkyō jinmei jiten* 二本佛教人名辭
典. Tokyo: Shinjinbutsu ōraisha, 1993. [NBJJ] {digitized by Charles Muller and Megumi Katahira}

Yi Chŏng 李政. *Hanguk pulgyo inmyŏng sajŏn* 韓國佛教人名辭典 (The Korean Buddhist Biographical Dictionary). Seoul: Pulgyo sidaesa, 1993. {digitized by Charles Muller and Asa Suzuki} [HPIS]

---

3. At the Chung-Hwa Institute for Buddhist Studies; Christian Wittern, et. al.

Bukkyō kanbon daijiten 佛教漢梵大辭典 (Buddhist Chinese-Sanskrit Dictionary). Hirakawa Akira 平川彰. Tokyo: The Reiyukai, 1997. [BCS]

Ding Fubao (Electronic version) -- [DFB]

Fo Kuang Shan Dictionary -- [FKS]

4. At the Research Institute of the Tripitaka Koreana; In-Sub Hur, et. al.

Korean readings for over 100,000 entries.

At the time of the most recent update, this compilation contained approximately 290,000 terms, and is now included within the full-text search function of the *Digital Dictionary of Buddhism*. If you have a CJK Buddhist lexical index that you would like to add, please contact Charles Muller at <acmuller@human.toyogakuen-u.ac.jp>. You can download this data collection.

Wherever possible, missing characters are encoded with Mojikyō numbers (&Mxxxxxx;). Kanjibase (&Cx-xxxx;) numbers that remain in the index have not yet been matched with their Mojikyō equivalent. Characters not yet contained in Mojikyō are encoded in algebraic format.