Chung-Hwa Buddhist Journal (2012, 25:105-128)
Taipei: Chung-Hwa Institute of Buddhist Studies
中華佛學學報第二十五期　頁 105-128 (民國一百零一年)，臺北：中華佛學研究所
ISSN:1017-7132

The XML-Based DDB:The DDB Document Structure and the P5 Dictionary Module; New Developments of DDB Interoperation and Access

Charles Muller (University of Tokyo)
Kiyonori Nagasaki (International Institute for Digital Humanities, General Incorporated Foundation)
Jean Soulat (Independent scholar)

Abstract

This paper has three parts. The first part, by A. Charles Muller,[1] consists of a comparative analysis of the DDB structure with that of the Dictionary Module in the current TEI P5 recommendations. The second and third parts are short summaries of the recent applications offering enhanced access to and usage of the DDB, created by Kiyonori Nagasaki[2] and Jean Soulat.[3]

Keywords:

Lexicons, XML, TEI, Web API, Interoperability

1 A. Charles Muller teaches Buddhism, East Asian thought, and a little bit of XML at the University of Tokyo. He is the founder and chief editor of the *Digital Dictionary of Buddhism*, its companion *Chinese-Japanese-Korean-Vietnamese/English Dictionary* (CJKV-E). He is also the founder and managing editor of the scholarly network H-Buddhism (<*http://www.h-net.org/~buddhism*>). His primary fields of research are Korean Buddhism and East Asian Yogācāra/Tathāgatagarbha thought, along with occasional forays into Zen, Confucianism, and Daoism. A listing of his books and articles on these topics can be accessed through his web site, Resources for East Asian Language and Thought (<*http://www.acmuller.net/index.html*>).

2 Kiyonori Nagasaki (永崎研宣) has an M.A. in Buddhist Studies from Tsukuba University, and is best known for his work as the primary developer handling the <SAT Taishō Database> and <INBUDS Database> in Tokyo. He has developed a range support structures to provide interoperation between SAT and the DDB, as well as INBUDS. He also wrote the Perl code for our "Feedback" option.

3 Jean Soulat is a telecom engineer with a personal interest in Buddhism and Chinese Culture. He has worked in the area of computer networks since the early days of the French public data network and then with different large scale networking and IT programs. He has created the application tool named Smarthanzi (<*http://www.smarthanzi.net*>) for looking up Sinitic words and characters in East Asian texts. Based on Smarthanzi, he has also created a specialized application for the DDB, called DDB Access <*http://download.smarthanzi.net/ddbaccess*>), which adds extensive functionality to the standard DDB lookup.

以可擴展標記語言(XML)為基礎的電子佛教字典(DDB): DDB文件結構與P5字典模組; DDB 相互操作與取用技術的新發展

Charles Muller (東京大學)
永崎研宣 (一般財団法人人文情報學研究所)
Jean Soulat (獨立學者)

摘要
此篇文章為三部分：第一部分由Charles Muller所寫，以現行的TEI P5所建議的字典模組，對DDB結構之比較分析；第二與第三部分則是由Kiyonori Nagasaki 與 Jean Soulat 所寫，簡短地概述近來提供加強對於DDB的取用技術與使用。

關鍵詞：詞典、XML、TEI、網路應用程式介面、相互操作

The DDB Document Structure and the P5 Dictionary Module

Charles Muller

No doubt that many of those of us who began their engagement in the development of web-based canonical collections, online databases, and various other research tools related to Buddhist Studies and East Asian studies at the time of the inception the WWWeb (circa 1994-95) look back in sheer amazement at the fact that almost fifteen years have passed since we made our most rudimentary stabs at developing these materials. At that time, Unicode, XML, Yahoo, Google, Internet Explorer, and scores of other now-commonplace Internet tools were yet to be heard of. In a short decade and a half, our way of doing research — and especially textual research — has been radically transformed.

Because of this radical change, young scholars coming into our field today need an entirely different set of skills for finding and organizing information. On the other hand, they no longer need, upon their departure from graduate school, to begin to try to figure out how they are going to afford to buy their first printed Taishō canon, and all the dictionaries and other reference tools needed to work with East Asian Buddhist texts. Most of these are now available digitally, and online, in one format or another. And these young scholars will have far more than simply the printed Taishō, Zokuzōkyō and other smaller canonical collections presently available at their disposal, as new, heretofore unavailable materials are being made searchable and downloadable — a main case in point being that of the newly developing Chan Texts Database, which will make available a variety of Chan texts, along with Dunhuang materials which were almost impossible to get one's hands on before this. And of course, for working with these texts, there is the DDB.

As has been explained over the years in numerous other presentations and project reports, I began my compilation of what turned out to be the DDB in my early days in graduate school (1986) having become aware of the incredible dearth in 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 in general. I worked at compiling terms for about ten years, and in 1995, shortly after the birth of the WWWeb, uploaded the collection that I had gathered up to that time up to my first web site, and the rest is history.[4]

4 For various accounts of the development of the DDB up to its present state, please see the bibliography, which provides a fairly complete listing of presentations, both published and unpublished.

Suffice it to say that the DDB has become the de facto choice among reference tools for young Western scholars doing work involving East Asian Buddhism. It is introduced as a primary reference work in all major North American universities that have programs dealing with East Asian Buddhism; it is supported in terms of content and programming by more than sixty scholars, many of whom are recognized as leading figures in their own sub-areas of Buddhist Studies or Information Technology; and it is presently subscribed to by twenty-eight university libraries.[5] It is also now accessible through online canonical text databases such as that of the SAT Taishō Database,[6] and is included in various Han-character-based lookup tools, including Smarthanzi,[7] the WWWebJDic Server,[8] and Tangorin.[9]

In prior papers dealing with the DDB, I have explained various aspects of the project, including history, design, collaboration strategies, XML structure,[10] and so forth. Here, I would like to focus on a specific issue with the present XML structure, paying special attention to its relation with the TEI P5 Dictionary Module.

At the conference where this paper was originally presented (which is the basis for the present volume), a significant portion of the presentations dealt with XML in one way or another. What most of them had in common, however, was their presentation of XML as a way of marking up pre-existent materials, whether they be pre-existent canonical collections, lexicons, or whatever. The DDB was unusual among the presented projects at this venue in that it was one of the very few where XML was shown as a framework for the development of a *new* data set from the ground up, and which, working through XSLT, provides the systematic structure for an online database-reference resource. Indeed, among online academic reference tools of its kind, the DDB as a fully XML structured resource is unusual, since most online reference resources tend to be run from a more traditional database structure.

The original choice of XML to structure the DDB data is basically an accident of history, related to the background of the people from whom I received my earliest technical advice. Most important in this regard is Christian Wittern, who discovered my earliest, hard-linked HTML version of the DDB on the Web sometime in 1995 or 1996. He applied a basic SGML structure to the data, where the tags referred to elements of the content and document structure, rather than being the mere style commands of HTML. Christian send me a copy of his SGML-restructured data, along with SoftQuad Panorama,

5 See <*http://www.buddhism-dict.net/ddb/subscribing_libraries.html.*>
6 See <*http://21dzk.l.u-tokyo.ac.jp/SAT/ddb-sat2.php.*>
7 Also available in the Windows desktop application DDB Access; both of these are available at <*http://www.smarthanzi.net*>; to be discussed by its developer below.
8 <*http://www.csse.monash.edu.au/~jwb/cgi-bin/wwwjdic.cgi?1C*>
9 <*http://tangorin.com*>
10 See JODI (2002).

a freeware viewer for SGML. I knew nothing of SGML at this time, but could see that it could be quite useful to mark up the data with content-meaningful tags as opposed to simple HTML style markers. Before long, the news of the impending release of the new XML standard to replace SGML had many people excited, and it seemed as if major software companies intended to support it, so I converted the DDB to XML, and stored it that way during the next couple of years, running an array of MS-Word macros to generate new HTML files periodically, uploading these to my web site.

But generating files this way each time was a convoluted and time-consuming process. Around 2000, I knew that people were beginning to publish XML materials on the web via XSL, and that more and more major markup and publication projects were turning to XML. But there were virtually no examples of serious real-world implementation other than in brief W3C explanatory materials. And without any kind of precedent available from which to learn, my programming skills were entirely insufficient for trying to implement raw XML on the web on my own.

At the same time, there were really no major data sets like my own readily available for testbed purposes, so newly appearing XML software development companies had no way to thoroughly test their new applications on actual large and complex data sets. On my website at the time, I had a description of the content and structure of my data (which was importantly, multilingual, including Chinese, Korean, and Japanese scripts, and a fairly large range of diacritical characters), and I was contacted by a few companies, including Microsoft and Altova (XMLSpy) who asked to use my data for testing of their XML software currently under development.[11]

After having been contacted by these companies, it occurred to me that there may be other individual XML developers who could take advantage of the DDB data for their own purposes, and at the same time help me to begin to take proper advantage of the XML structure and begin delivering the data on the web in real time, through XSL and whatever else might be required. I posted a message on the Mulberry XSL list inquiring as to whether anyone was interested in working with my data in this way. Within a week, I received a response from Michael Beddow, who, in connection with work he was doing on the web-based version of the Anglo-Norman Dictionary,[12] expressed a willingness to try to get the XML-DDB up and running on a server. In a very short time, he accomplished this to a level far beyond what I could have ever hoped. Since both Michael

11 I agreed to both of these requests. From Microsoft, I never even received so much as a thank-you note. Altova gave me one free upgrade (to my already-purchased license), but then forgot about me, demanding that I pay the full price for their next Enterprise version. This turned out to be a major motivation toward my efforts to learn Emacs. Also, luckily, not too much later, <oXygen/> made its appearance on the scene, with its much more reasonably priced, fully-featured XML editor, and much more humane support staff.

12 <*http://www.anglo-norman.net/*>

and I have recounted the main points of this landmark task in some detail in the past,[13] I will not go into great detail retelling this stage of the process, except to say that Michael is still providing the basic technical support for the project, including security as well as the basic delivery of the data, for which I, and thousands of researchers of Buddhism around the world can be eternally grateful.

This simple but elegant XML/Perl/XSL delivery system developed by Michael has functioned in the same way, basically unchanged for almost a decade, and technically speaking, there have been no special demands or changes to our system that XML/XSL can't deal with, so although the suggestion of changing over to a traditional database system has been made to me from time to time, I have never felt the need to give it serious consideration. Although a database setup such as MySQL may be a bit faster in retrieving entries, having the data in XML format allows me to fully integrate it with the rest of my work on my desktop. Since I do virtually all of my scholarly research and translation in XML, and maintain various related data sets in XML or plain text format, having the DDB in XML while using the same basic tag structure for the rest of my documents makes it very easy to move things back and forth.

Having mentioned the fact that I use the same basic tagging structure in the rest of my work, I would like, from here, to deal with a technical aspect of the project that I have touched on briefly from time to time, but have never really worked through in detail: that is the relationship of the structure of the DDB to the TEI document model. I have been using TEI for my writing and most of the other phases of my work for about eight years now. Also, the two major technical contributors to the DDB project, Christian Wittern and Michael Beddow, are persons well-versed in the development and implementation of TEI. Since TEI has a subset of tags specifically designed for the structuring of lexical materials, it might be reasonable to assume that the DDB would be a fully TEI-based project.

It is to a significant degree. Since I have been using TEI in my work for several years, it has been the case that when I have needed a new tagging structure for the DDB, I have always first checked the TEI tag set to search for an appropriate tag. Almost always finding one, I have done my best to implement new elements in the DDB according to TEI hierarchical rules and with the recommended attributes. Thus, the content of the <sense> nodes in the DDB (discussed in further detail below) is fully TEI(P4)-compliant. This covers many sub-structures, including <list>, <biblStruct> and many other basic prose structures necessary for writing short dictionary entries, as well as encyclopedic entries—basically replicating the rules of what would be allowable inside the TEI <p> element.

13 See the JODI (2002) article, ibid. I have also discussed Michael's role in the project in a few other articles.

For the nodes above, and outside <sense> however, the structure of DDB entries is somewhat different from the sort of thing that one would build if one were to start from the ground up with the present TEI P5 Dictionary Module. When the XML for the DDB was first set up, there was no special intention to reject the TEI (at that time P4) structure. Christian Wittern and I sat down at a conference one time and tried to write a tag structure that best fit that of the DDB at the time. At this time I knew nothing of TEI, and Christian was just getting seriously involved in this Initiative. Thus, while this initial structure was informed by TEI concepts, it tended to conform more closely to the actual structure of the DDB, rather than trying to force a full TEI framework.

The basic structure of a DDB entry is currently like this:

<entry> (one dictionary entry)
<hdwd> (Chinese logographic headword)
<pron_list> (grouping the pronunciations into a separate node)
<pron> (pronunciations of the headword in various East Asian languages, in roman script as well as native syllabaries)
<pron>
<pron>
. . .
</pron_list>
<sense_area> (grouping semantic/content information)
<trans> (a short, primary translation or meaning of the head word)
<sense> (explanatory portion of the headword, for which there is usually more than one)
<sense>
. . .
</sense_area>
<dictref> (list of references to entries for the term in other major reference works)
<dict>
<dict>
. . .
</dictref>
</entry>

Filled out with attributes and data, a relatively short sample entry looks like this:

<entry ID="b9403" added_by="cmuller" add_date="1993-09-01"

update="2009-11-25" rad="金" radval="08" radno="167" strokes="12">
<hdwd>鐃</hdwd>
<pron_list>
<pron lang="zh" system="py" resp="c.wittern">naó</pron>
<pron lang="zh" system="wg" resp="cmuller">jao</pron>
<pron lang="ko" system="hg" resp="cmuller">요</pron>
<pron lang="ko" system="mc" resp="cmuller">yo</pron>
<pron lang="ko" system="mr" resp="cmuller">yo</pron>
<pron lang="ja" system="kk" resp="cmuller">ドウ</pron>
<pron lang="ja" system="hb" resp="cmuller">nyō</pron>
<pron lang="vi" system="qn" resp="daouyen">nao</pron>
</pron_list>
<sense_area>
<trans resp="cmuller" rend="hide">a <term lang="en">hand-bell</term></trans>
<sense resp="cmuller" ref="Yokoi,Hirakawa">Cymbals.(Skt. <term lang="sa-mw" n="11740">tūrya</term>) <bibl type="canonlink">法華經 <xref canonref="http://21dzk.l.u-tokyo.ac.jp/SAT/T0262_,09,0009a11:0262_,09,0009b11.html">T 262.9.9a13</xref></bibl> </sense>
</sense_area>
<dictref>
<dict><title>Zengaku daijiten (Komazawa U.)</title><page>989b</page></dict>
<dict><title>Japanese-English Zen Buddhist Dictionary (Yokoi)</title><page>512</page></dict>
<dict><title>Ding Fubao</title><page/></dict>
<dict><title>Buddhist Chinese-Sanskrit Dictionary (Hirakawa)</title><page>1193</page></dict>
<dict><title>Bukkyō daijiten (Mochizuki)</title><page>(v.1-6)1307b,2595b,4137a</page></dict>
<dict><title>Bukkyō daijiten (Oda)</title><page>1370-1</page></dict>
</dictref>
</entry>

We will get into the treatment of comparative issues in detail below, but just to provide the reader with some context, it is probably useful to have some idea of the basic P5 recommendation for dictionary structures, which, as provided on the TEI web site,[14] is like this:

```
<entry>
<form>
<orth>disproof</orth>
<pron>dis"pru:f</pron>
</form>
<gramGrp>
<pos>n</pos>
</gramGrp>
<sense n="1">
<def>facts that disprove something.</def>
</sense>
<sense n="2">
<def>the act of disproving.</def>
</sense>
</entry>
```

As we can see, the fundamental elements <entry>, <pron>, and <sense> are used in the DDB for the same purposes, and with basically the same kind of hierarchical structure. The most glaring difference is seen in the DDB element <hdwd> (head word), which is idiosyncratic, an odd tag that I created during a short period in which the DDB was stored in a mixed structure of XML and HTML tags, and the attempt to use the tag <head> for head words produced obvious problems in HTML. It would have been better to get rid of this at an earlier stage, but opportunities were missed, so it remains here, embedded at the core of the DDB. In P5, the corresponding tag would probably be <orth>. Beyond this, the other major difference is the presence in the DDB of the node <dictrefs>, within which information is contained regarding references on the same term in other dictionaries.

Before we address specific issues of tags and their structure, a word regarding the nature of the data set itself is in order. That is, the East Asian notion that is translated into English as "dictionary" — 辭典 (Ch. *cidian*; K. *sajeon*; J. *jiten*) sometimes refers to

14 <*http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-entry.html*>

something that is basically the equivalent of a Western dictionary. But it is also often something quite different, in the sense that it may well end up containing entries that are more like those of an encyclopedia in terms of length and complexity. And there is less linguistic-oriented information (such as grammatical forms and so forth).

Another distinctive characteristic for East Asian works of this type is that with Sinitic Buddhism being a pan-East Asian phenomenon, the Chinese logographic head words have distinct pronunciations in Mandarin, Korean, Japanese, and Vietnamese (including variant readings within these languages), with these being represented in both native syllabaries and Western romanization systems. Since the TEI dictionary module is basically constructed upon the Western model, problems will be evident from the start. Acknowledging these points, let us try to see what would be involved in bringing the DDB structure in line with TEI P5. For the moment, we will leave the level of attributes aside, focusing on elements.

The first obvious change would be that of replacing <hdwd> to <orth>. This would not be terribly difficult, since a global replacement throughout the data and XSL files should not result in any major problems. Next, removing the <pron_list> wrapper from the level above the <pron> elements would not pose major problems at the XML level, but it would require some degree of rewriting of the style sheet; the same would be true for removing the <sense_area> wrapper from around the senses. The TEI element <gramGrp> is not relevant to the DDB.

A major consideration would be the conversion of the <dictrefs> area. This is an idiosyncratic component of the DDB, since it is not customary for dictionaries in general — whether they be Eastern or Western — to provide a list of references in other dictionaries or encyclopedias. Among the child nodes offered in TEI entry/dictionaries, the only thing that appears to come close to this is the element <xr> (x-reference). But if we used this, it would probably be more appropriate to use it in the place of the <dict> reference, rather than the wrapper <dictrefs>.

With this kind of list, including essentially bibliographical references, it would be helpful for styling and other programming purposes to have a wrapper for this list of <xr> elements, something playing a similar to <listBibl>. Actually, the inclusion of <listBibl> — or something like it — as a possible child of <entry> would very helpful in this case. Then would could convert the <dictrefs/dict> structure into a basic TEI <listBibl/bibl> tree (<listBibl> does appear under <xr>, so this would be another possible route.). But again, this is a special dimension of the DDB, and not a something that one would see needed for dictionary entries in general.[15]

15 In this regard, the DDB is often more like an encyclopedia than a dictionary, but the TEI does not at the moment have an encyclopedia module. A discussion of encyclopedias on TEI-L (<*http://listserv.brown.edu/archives/cgi-bin/wa?A0=TEI-L*>) at the end of 2009 concluded in

A provisional rewrite of the DDB entry structure, based on the above changes, would now look something like this (shortening some sections for the sake of readability):

```
<entry ID="b9403" added_by="cmuller" add_date="1993-09-01" update="2009-11-25"
rad="金" radval="08" radno="167" strokes="12">
<form>
<orth>鐃</orth>
<pron lang="zh" system="py" resp="c.wittern">naó</pron>
<pron lang="ko" system="hg" resp="cmuller">요</pron>
<pron lang="ja" system="kk" resp="cmuller">ドウ</pron>
</form>
<sense type="brief"><def>a <term lang="en">hand-bell</term></def></sense>
<sense type="normal" resp="cmuller" ref="Yokoi,Hirakawa">Cymbals.(Skt. <term
lang="sa-mw" n="11740">tūrya</term>) <bibl type="canonlink">法華經<xref
canonref="http://21dzk.l.u-
tokyo.ac.jp/SAT/T0262_,09,0009a11:0262_,09,0009b11.html">T
262.9.9a13</xref></bibl> </sense>
<xr>
<bibl><title>Buddhist Chinese-Sanskrit Dictionary (Hirakawa)</title><biblScope
type="pages">1193</biblScope></bibl>
<bibl><title>Bukkyō daijiten (Mochizuki)</title><biblScope type="pages">(v.1-
6)1307b,2595b,4137a</biblScope></bibl>
<bibl><title>Bukkyō daijiten (Oda)</title><biblScope type="pages">1370-
1</biblScope></bibl>
</xr>
</entry>
```

The next level of conversion—that of attributes—gets more complicated, as the DDB utilizes a number of attributes that are not contained either as attributes or elements in the dictionary module or elsewhere in the TEI P5 tag set, as far as I can determine. The character of the attributes currently used in the DDB can serve to draw our attention to

---

the recommendation for the encyclopedia maker to structure his data with a series of nested tags with various attributes making the content distinctions.

some of the distinctive aspects of the DDB mentioned above. That is, rather than being the markup of some pre-existent lexicon, the DDB is a new work in progress. To properly embed information related to the development of each entry, the attributes attached to our <entry> tag contain several pieces of information that provide important history regarding the entry, as well as categorizing and sorting information. These include, at the entry level, @added_by, @add_date, and @updated. Interestingly, the TEI has always shown concern about this kind of documentation, as these kinds of elements have always been part of TEI document headers. But as far as I can tell, there is no mechanism for recording this kind of information at the level of entries or entry child nodes in a reference work. So if we tried to convert to P5, these would need to be added to a customized schema. Similarly, at the <sense> level, the @resp, @source, and @ref attributes are critical to the DDB for keeping clear records of sources, contributions, responsibility, and related references. Unless I have missed some alternative way of dealing with these in the Guidelines, it seems that the committee that developed the dictionary module had in mind the markup of pre-existent dictionaries, rather than the collaboration-based creation of a new dictionary in mind when they created this attribute structure.

Would it be worth the effort to convert to P5? The thought of going through this present comparison of the DDB entry structure with that of P5 has been on my mind for some time. Why would one go through the trouble of making this kind of major conversion in an XML structure that is working fine as it is?

There would be a few significant advantages to doing this. The first reason is that, as mentioned above, most of the rest of the academic research and writing that I am doing is being composed in TEI P5. Having a DDB structure that is fully TEI compatible would allow me to freely copy data back and forth without generating non-validity problems at either end. Second, this would allow the usage of the same basic style sheets for all of my projects. Third, full TEI compatibility would allow me to take advantage of other tools produced by members of the TEI community, including its schemas, and CSS/XSL sheets.

There are, however, a couple of significant drawbacks. First, it would not only require a major reworking of the data and the style sheets. It would also entail a reworking of scores of MS-Word macros that have been the background for the actual production of the data for more than a decade. So careful consideration is needed before taking the leap.

Interoperation I: The DDB and SAT

Kiyonori Nagasaki

The digitization of the resources for Buddhist studies—as well as those for other fields of academic inquiry—has now been in progress for a few decades. As a result of the diligent efforts of those engaged in various digitization projects, researchers of Buddhism now have access to a wide range of electronic materials, a state of affairs that serves to enhance the efficiency, accuracy, and overall quality of their research. The emergence of the Web environment has been the fundamental catalyst allowing a wide range of new ways of storing, representing, and sharing of resources. Recently, the next evolution of the Web—known as Web 2.0—has brought about a transformation in the delivery and handling of digital scholarly resources for all kinds of research. Most important here is the availability of the AJAX technology and Web API, which have enhanced the ways of sharing and delivering Web resources by leaps and bounds. The dissemination of cloud computing technology will further serve to support these kinds of developments.

Even only a decade or so ago it was taken for granted that for complementary digital resources—such as text and lexicon—to work together effectively, they had to be integrated one way or another into a single database format. While this may still well probably happen in a case where both resources are developed by the same individual or within a single project, if the resources were developed by separate entities, the combining of both into a single structure would usually entail the loss of independence or identity for one party or the other. However, in recent years, the situation has changed significantly, since, by adopting AJAX, Web API, and similar technologies, those who have been developing Web-based resources in the Humanities will be able to cooperate/interoperate between projects while each project maintains full independence.

The prominent example to be offered here is the recent interoperation developed (starting in 2008) between the SAT Taishō Database[16] and the DDB and INBUDS (Indian and Buddhist Studies Treatise Database)[17] on the Web environment using the AJAX technology. Since April of 2008, the SAT Web service has been providing the function wherein if the user selects a portion of *kanbun* text from the Taishō canon with the mouse, a list of terms within that text that are available in the DDB will be generated alongside the text, along with English head words and links into the DDB itself. We are continuing to enhance various aspects of this function.

Since the time of the presentation of this application at the Chan texts conference in Oslo in October 2009, SAT has been providing further new functions implemented with

16 <*http://21dzk.l.u-tokyo.ac.jp/SAT/search.php*>
17 <*http://21dzk.l.u-tokyo.ac.jp/INBUDS/search.php*>

AJAX and Web API. Previously, users could search related articles from the INBUDS database (maintained by the Japanese Association of Indian and Buddhist Studies[18]), but were only able to elicit basic bibliographical reference information. Under this new function, users are also able to obtain PDF files of the articles (when PDFs are available) by clicking on the PDF icon displayed on the ending of each line of the search results. Clicking on the icon opens up a page within the CiNii service[19] that includes a link to the PDF file. This PDF file service is provided for the whole academic society, not only to Buddhist Studies or the Humanities. CiNii distributes their bibliographic data as a PDF file through their Web API service.

INBUDS has taken optimal advantage of this public service by providing a Web API that allows other Web services to retrieve the INBUDS search results, including their PDF file information. The SAT Web service has implemented this, but it is important to know that every scholarly web service is welcome to enrich itself by taking advantage of CiNii's offering. Furthermore, the SAT Web service has been further contributing to CiNii's efforts by providing some Web APIs. SAT is also planning to provide some more efficient APIs so that the other Buddhist service providers can also distribute better services. Adopting AJAX and Web API, each project/service can enrich other services, while maintaining their independence as individual projects.

In this kind of Web environment, we will have the opportunity to work together not only as isolated contributors of data but also as individual and cooperative service providers so that researchers in our field can benefit from more efficient services. By so doing, our study and inner space will be greatly enriched.

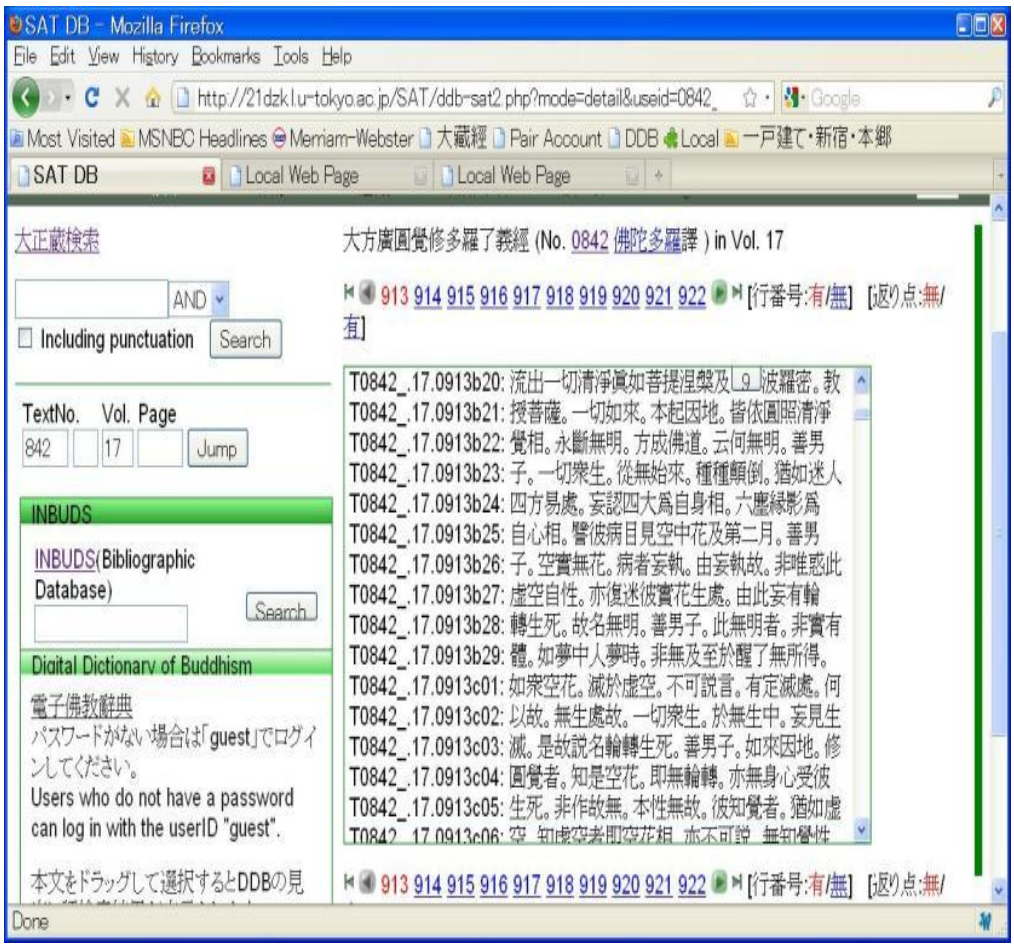
18 <*http://www.jaibs.jp*>

19 (Scholarly and Academic Information Navigator, pronounced like "sigh-knee") is a database service maintained as a Japanese government project by the National Institute of Informatics, which enables searching of information on academic articles published in academic society journals or university research bulletins, or articles included in the National Diet Library's Japanese Periodicals Index Database. <*http://ci.nii.ac.jp/en*>

1. DDB Parsing From the SAT Database

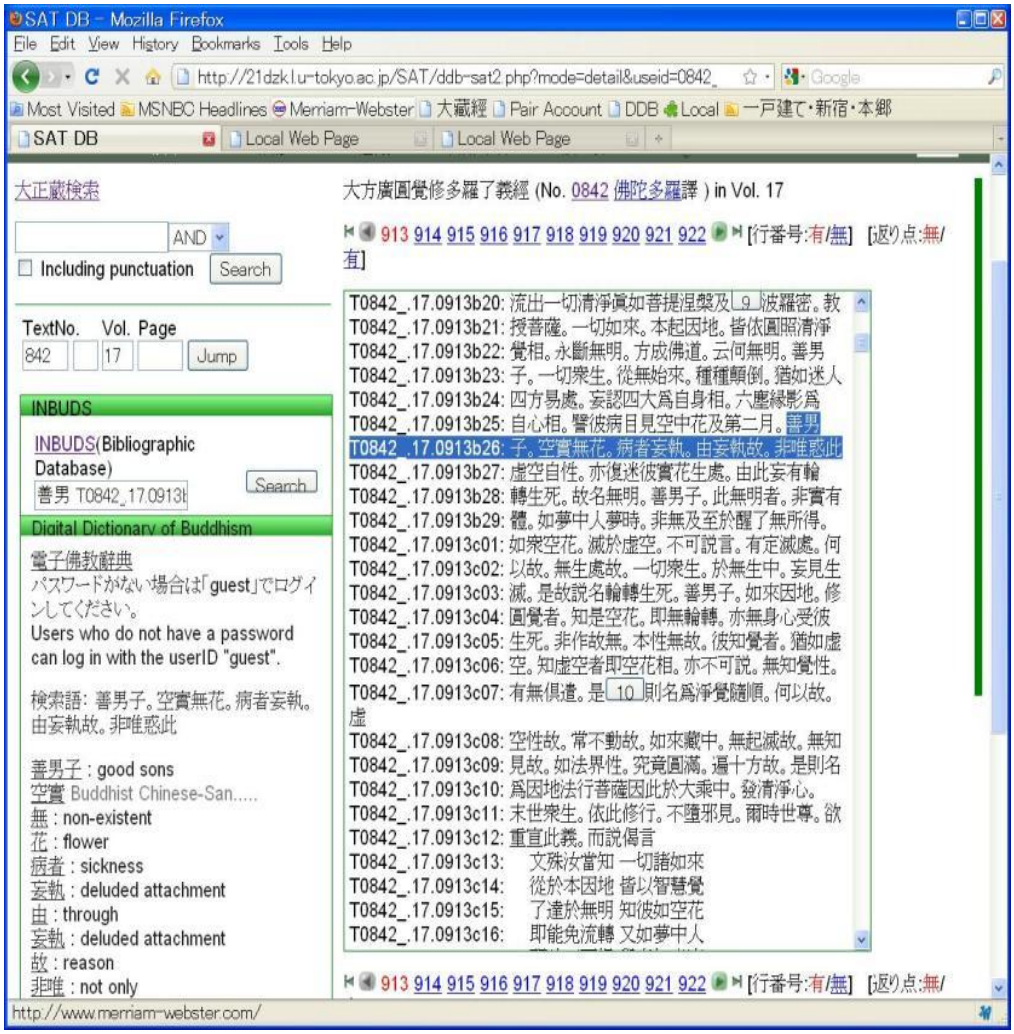
1.1. SAT Text View

The user opens up desired text by scrolling or computer search:



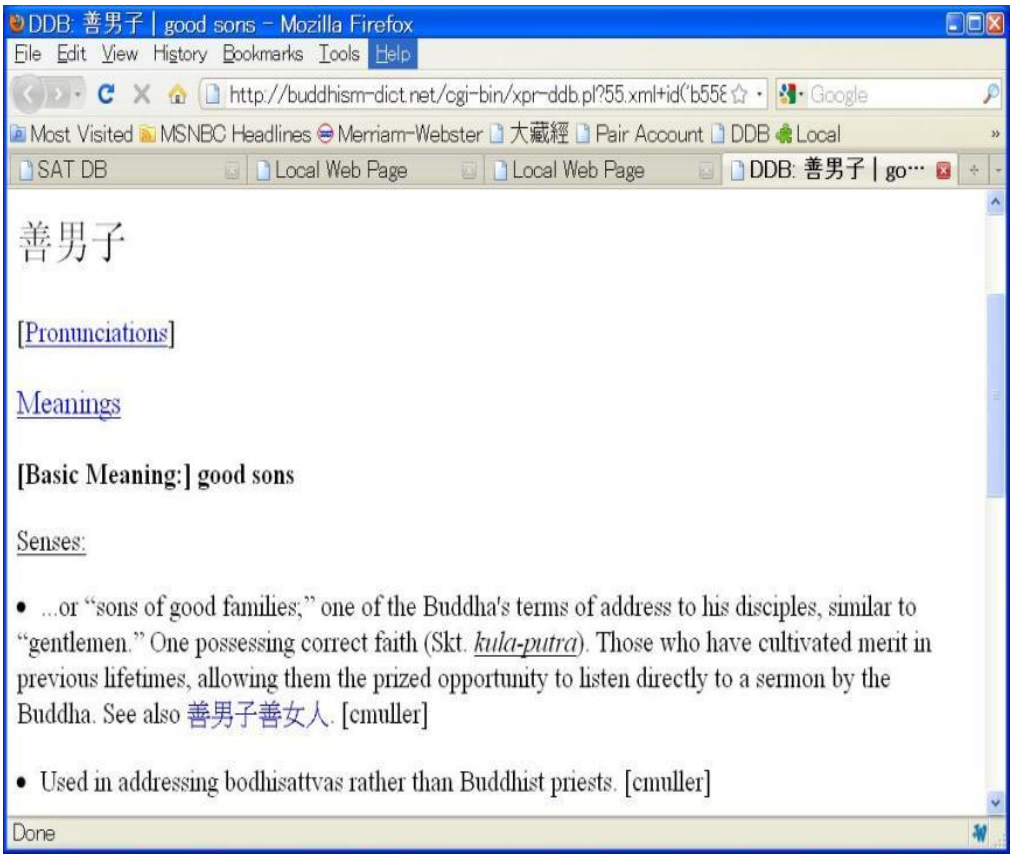
1.2. Selecting and Generating a Word List

One then selects a portion of text with the mouse, upon which the DDB words contained in the selected text will be arranged in a list on the left:



1.3. Lookup in the DDB

Clicking on term in the list will open up the entry in the DDB:



Interoperation II: DDB Parsing and Lookup with SmartHanzi.net and DDB Access

Jean Soulat

1. Smarthanzi.net

SmartHanzi.net is a website with a parsing and lookup tool developed by Jean Soulat for Chinese and links to etymological lessons by Dr. L. Wieger, S.J., in his *CHINESE CHARACTERS – Their origin, etymology, history, classification and signification*[20] Finding a given character in this book can be a trying experience, since one is forced to work through a number of indexes. The introduction of simplified characters in continental China has added a further level of complication. This is precisely the sort of situation where information technology can be of the greatest help: with just a mouse click, the website points to the relevant etymological lesson (out of 177) and phonetic series (out of 858).

1.1. Parsing and Lookup

Parsing and lookup relies on various Chinese word lists available on the Internet:

- For basic Chinese, CEDICT MDBG (English), HanDeDict (German); for Buddhism, the DDB and Soothill & Hodous.
- The companion site Smartkanji.net uses the JMDict multilingual list for Japanese available on Jim Breen's Monash Nihongo FTP Archive.[21] Jim Breen also kindly provides Japanese specific tables for adjectives and words.

When a text is submitted to the application, the server parses it and displays a first view of all words found (in the main list) just under the text. Users can then lookup anywhere in the text either with a mouse click or by dragging over if more convenient. The website shows all words recognized at the mouse position. It does *not* try to make a choice.

Users have to select one among the available dictionaries. If one needs to lookup from several dictionaries, several tabs in the browser window offer a convenient solution.

20 First published in 1899 (French) and 1915 (English), based on the 2nd century *Shuowen Jiezi*. This work contains numerous technical errors, but is a valuable historical document in that it reflects the understanding that many Chinese had regarding their writing system.

21 <*http://www.bcit-broadcast.com/monash/*>

1.2. Technology

SmartHanzi.net uses the so-called "Ajax" technology: one HTML page is used as an application. Further data are then updated through XmlHttpRequest and JavaScript within the original HTML page. The server is written in PHP and uses flat files (no database) to keep parsing time acceptable.

Since some large size tables need to be loaded for each text, the website works best when users submit full paragraphs or short texts.

1.3. Limitations

The word lists available on the Internet are convenient for parsing and lookup. But they do not contain enough detail to navigate from one word to another, as many people love to do with paper dictionaries. This is where the DDB XML access provides a great opportunity.

1.4. The DDB Access Application

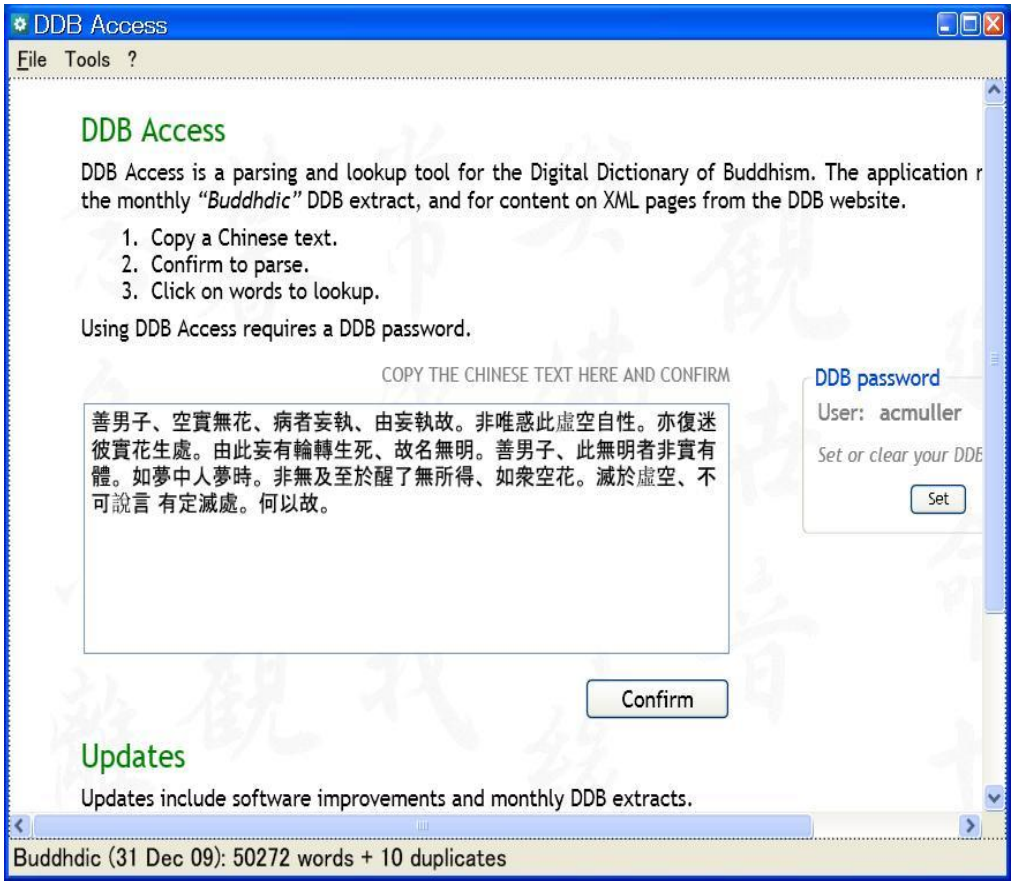
Both Smarthanzi (<*http://www.smarthanzi.net*>) and DDB Access (<*http://download.smarthanzi.net/ddbaccess*>) work off of the same DDB data extract, which includes headwords, pronunciations, and basic definitions in a public file published monthly by Charles Muller.[22] It does not include the full set of data accessible through the DDB website (<*http://buddhism-dict.net/ddb*>). Since the full data displayed on the DDB website are also available in XML format on per word requests, there was a perspective to put together the parsing and lookup facility of SmartHanzi.net and the complete DDB data.

22 This file, <*http://acmuller.net/download/buddhdic.txt.gz*> is the same as that which is published on Jim Breen's WWWJDic Server (<*http://www.csse.monash.edu.au/~jwb/cgi-bin/wwwjdic.cgi?1C*>). It is also used by the SAT database to access terms in the DDB, as well as by the developer of Tangorin.

1.5. Views of DDB Access

1.5.1. Step One: Paste in Text

The user pastes in some East Asian text containing Chinese characters:

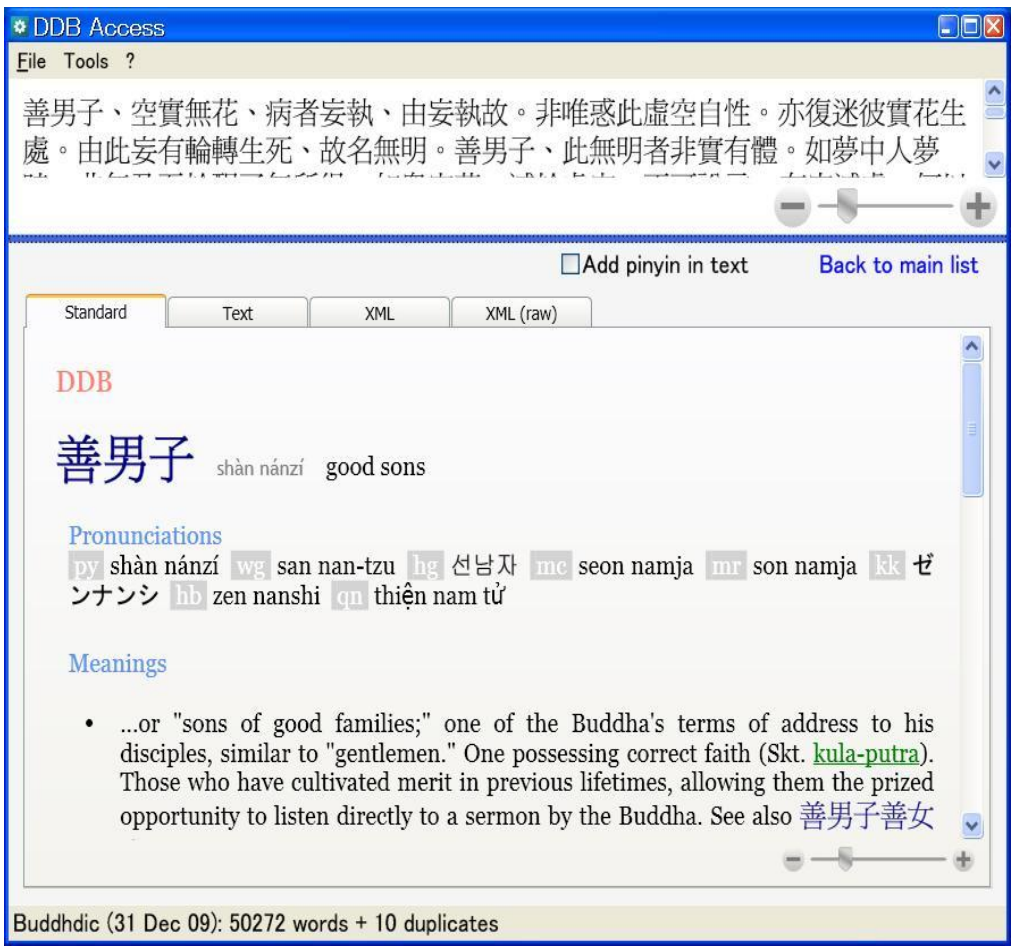


1.5.2. Step Two: Parse Text

The text is then parsed, separating compound words on the left and single characters on the right:

1.5.3. Step Three: Select Word for Lookup

The user can then select a character or compound word from the generated list for lookup in the DDB:



1.6. Meeting the DDB Password Policy

Both for security purposes, and in order to encourage users to contribute to the DDB, Muller has implemented a tiered password policy. This has led to a maximum number of calls per day for users with the "guest" login. In order to meet the DDB password policy, DDB data have to be requested from the user PC. Since making XML calls from the SmartHanzi.net server would have infringed the DDB password policy it was not considered.

One option might have been to include XML data into SmartHanzi.net HTML / Ajax pages, of course subject to agreement from the DDB team. However, for security reasons, JavaScript does not allow a web page from one site (SmartHanzi.net) to access XML data from another site (DDB). Cross-domain calls may be allowed in latest generation browsers but not all users have a recent browser.

1.7. The DDB Access Application

The chosen solution was to develop a PC application, called "DDB Access," which is not subject to the cross-domain limitation. The application has the same look and feel as the website and uses the same DDB extract to parse the text. When the user clicks on a word, the application makes a request to the DDB server and gets the full XML data. Each user needs to provide his or her DDB password.

The XML data are presented with different views in separate tabs:

- Standard view: similar to the DDB website.
- Text view: no formatting, convenient for paste and copy into a word processor.
- XML: formatted XML view.
- XML (raw): unformatted XML view, for paste and copy into a XML editor.

To make sure that parsing and lookup maintain consistency, it is recommended to download monthly updates.

1.8. Technology

The application is developed with Microsoft Windows Presentation Foundation (WPF, .NET 3.5). It embeds a SQLite lightweight database which makes it easy, for instance, to add the "Also contained in" function.

1.9. Soothill and Hodous

Both SmartHanzi.net and DDB Access also include Soothill and Hodous entries, as digitized and published by A. Charles Muller. (<*http://www.acmuller.net/soothill /index. html*>).

References

Muller, A. Charles. 2009. *The Digital Dictionary of Buddhism [DDB] as a Model for Web Collaboration*. Symposium of the Information Processing Society of Japan, University of Tokyo.

----. 2009. The Digital Dictionary of Buddhism [DDB]: Present Status and Future Developments. *Scholars of Buddhism in Japan: Buddhist Studies in the 21st Century.* Kyoto: International Research Center for Japanese Studies. 87–100. *http://acmuller.net /articles/ddb-nichibunken-200803.html*.

----. 2005. *A Model for Scholarly Collaboration in the Development of On-line Reference Works: The Digital Dictionary of Buddhism.* Conference on New Technology in the Handling of East Asian Documents; Chinese National Library, Beijing. *http://acmuller. net/articles/ddb-beijing-conference.pdf*.

----. 2002. Moving into XML Functionality: The Combined Digital Dictionaries of Buddhism and East Asian Literary Terms. *Journal of Digital Information: Special Issue on Chinese Collections in the Digital Library* 3(2). *http://journals.tdl.org/jodi article/view/83/82*.

----. 2001. ディジタル媒体を使用して、仏教データの調査と普及: 仏教学のディジタル辞書. (*Using the Digital Medium for Research and the Dissemination of Buddhist Studies Data: The Digital Dictionary of Buddhism*). Annual Conference of the Japanese Association for Indian and Buddhist Studies, Tokyo University. *http://acmuller. net/articles/jaibs2001.html*.

----. 1999. *Developments of the Web Dictionaries of East Asian Thought: Stepping up to XML*. Seminar on Computing in East Asian Studies, Kyoto University Computing Center.

----. 1999. *Update on the Development of the Digital Dictionary of East Asian Buddhist Terms*. Fifth meeting of the EBTI, Academia Sinica, Taipei. *http://acmuller.net/articles report1999ebti.htm*.

----. 1998. *The Structure and Function of the Interlinked Electronic CJK-English and Buddhist CJK-English Dictionaries*. International Conference of Asian Scholars (ICAS), Leiden University. *http://www.acmuller.net/articles/dictionaries1.htm*.

----. 1998. *The Structure and Function of the Interlinked Electronic CJK-English and Buddhist CJK-English Dictionaries.* Meeting of the Pacific Neighborhood Consortium (PNC), Taiwan.

----. 1997. *The Usage and Development of Digital Reference Tools in Working with CJK Buddhist Texts: Interlinked CJK and Buddhist CJK Dictionaries*. Fourth Meeting of the Electronic Buddhist Text Initiative, Ōtani University, Japan.

----. 1996. *Introducing the Web Dictionary of East Asian Buddhist Terms*. Third Meeting of the Electronic Buddhist Text Initiative, Foguang Shan, Taiwan.

Nagasaki, Kiyonori; Muller, A. Charles; Shimoda, Masahiro. 2009. Aspects of the Interoperability in the Digital Humanities. *Digital Humanities.* 375-377.