A Model for Scholarly Collaboration in the Development of On-line Reference Works: The Digital Dictionary of Buddhism

Charles Muller
Toyo Gakuen University
Kyoto University Institute of Humanities Presentation
Beijing January 22, 2004

I. Technical Review

I began the compilation of the Digital Dictionary of Buddhism (DDB) and the CJKV-English Dictionary soon after my entry into graduate school in Buddhist Studies, upon my coming to awareness 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 in general. I decided, during my first Buddhist and Confucian/Daoist texts readings courses to save everything I looked up, and have continued that practice down to the present, through the course of studying scores of classical texts.

At the time that I began this process, I could not have dreamt 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 new, comprehensive 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. Soon after this, the dictionary was discovered on by

Christian Wittern, who promptly downloaded all the files, and applied a basic SGML structure, which is the ancestor of the XML markup system used today.

Due to the lack of widespread popular implementation of SGML, I did not make any special effort to develop this format for a few years. But after 2000, the popularity of XML began to suddenly increase, and so I began to take this format seriously. A major technical turning point in the history of the DDB came in January 2001, when I was contacted by Michael Beddow, a scholar of German Literature who was also an extremely accomplished XML programmer, and who had been using XML for some time already to develop his own only lexicographical project, the Anglo-Norman Dictionary (*http://www.anglo-norman.net*). Michael generated, based on the markup structure of the DDB, an array of indexes that used Xpointers to call up single-entry data units out of large files, each of which contained hundreds of entries. Michael also developed a CJK-Utf-8 search engine.[1]

II. Content Development

In my first presentation of the DDB at the meeting of the Electronic Buddhist Text Initiative (EBTI) in 1996, the dictionary had 3,200 entries. Today, less than nine years later, the DDB now boasts 35,000 entries, making it by far the largest compilation of its type in the English language, and even larger than some of the best-known Japanese works, such as Oda's *Bukkyō daijiten*. An instrumental factor in this rate of growth is the aid received through JSPS research grants, which allowed us to hire graduate students to help digitize large amounts of data for input. But this stage ended in 2002, and we have entered a new phase, where we are finally receiving large contributions of data from

1 I have focused here on developments in the DDB, but please note that all of the same technological enhancements have been applied to the CJKV-E.

unselfish collaborators who understand the spirit of the project and its limitless potential for the future. Two of the largest recent individual contributions have come from Prof. KARASHIMA Seishi, who has contributed over 7,000 entries from his research on the *Lotus Sutra*, and from Dr. Stephen Hodge, who has contributed over 2,500 hundred terms from his translation work on the *Yogācārabhūmi-śāstra.* In addition to these unusually large contributions, we have recently been benefiting from a continuous stream of smaller contributions, amendments, and corrections, from an ever-increasing number of scholars.[2]

While the DDB can certainly be viewed as a fairly successful model of the possibilities of online collaboration, it should be made clear that until we set up a mechanism to strongly encourage (perhaps "force" is the better term here) contribution, voluntary data submissions were few and far between. Initially we set up our password access/quota system to deal with hacking and data-theft problems. But we also discovered that we could take advantage of this same system to encourage contributions. Through this system, users who log onto the DDB web site to search for terms are able to freely look up ten items in a 24-hour period. After this, they are greeted by a message telling them that their quota is finished, but that they may gain an unlimited quota password by making a small data contribution.

In earlier days, when the content coverage of the DDB was still rather limited, this strategy did not generate that much response. But during the past year, with the expansion of the coverage to its present number, usage of the resource has also increased. The DDB has become a standard lookup tool for many Buddhist studies specialists—especially those who are doing intensive translation work. It is also used extensively in university classes in North America, and is a basic research tool listed on the syllabi of Buddhist

2 For a full list of contributors, see *http://www.acmuller.net/credits/credits-ddb.htm*.

Studies courses in such prestigious institutions as Harvard, Stanford, Princeton, Columbia, Berkeley, and other universities. As the DDB grows in both usefulness and in reputation as an essential reference tool for Buddhist studies research, scholars and students are increasingly coming to depend upon it, and thus eventually come to need unlimited access. Most serious scholars already have a large amount of specialized information on their hard drives that can easily be modified to become a DDB entry. All they need, it seems, is a small reason to make this effort, along with a little prodding.

For interested persons who do not have the specialized training to write or edit DDB entries, paid subscriptions are available. This approach was settled upon not with the expectation of making a lot of money, but simply to provide a recourse for persons who demanded full access in one way or another. As a by-product of this offering however, we decided to offer institutional subscriptions as well, and recently a number of major universities have decided to have their libraries subscribe to the DDB, including Columbia, Berkeley, Santa Barbara, and UCLA. While we are happy to gain a small amount of money to put back into the project, at this point, the greatest value of these subscriptions is in the recognition being accorded to the DDB as a primary reference tool. It is especially significant that this reference tool has been put together and produced, not by a major publishing company, but by a group of like-minded scholars.

III. The Structure of a DDB Entry

As mentioned earlier, the DDB uses XML as its basic structural format. The DTD is based loosely on the recommendations of the Text Encoding Initiative, using many of the entities and attributes that are used for lexicons.[3] An entry is divided into three major

3 The reason that the DDB is not based more fully on TEI is simply that most of the structure was developed before I adequately understood the TEI model. I have thought from time to time about redoing the whole structure according to the TEI DTD, but the retooling of the stylesheets, as well as

sections: (1) A Pronunciation Section, wherein the readings of a Chinese term are provided in various East Asian languages and their romanization systems. (2) A Sense Section, which provides the translation of the term and other explanatory material, and (3) An External References Section, which provides references to the term in a variety of Buddhist Studies reference works. Each of these larger nodes has children nodes, and various other entities contained within. When a user selects a term either via hyperlink or by search engine lookup, and HTML page is generated. One sample page is given below:

numerous other aspects of the web implementation of the data set are simply too daunting for me to seriously consider at this point in time.

言說

[Pronunciations]

[py] yánshuō
[wg] yen-shuo
[hg] 언설
[mc] eonseol
[mr] ŏnsŏl
[kk] ゴンゼツ
[hb] gonzetsu

Meanings

[Basic Meaning:] verbal expression [s.hodge]

Senses:

- (Skt. *vyavahāra*; Tib. *tha snyad*) [s.hodge]
- expresses, recounts; (Skt. *abhi-lap**; Tib. *brjod par 'gyur ba*) [s.hodge]
- expressing, an expression, an utterance; (Skt. *abhilāpa*; Tib. *brjod pa*) [s.hodge]
- discourse; (Skt. *kathā*; Tib. *gtam*) [s.hodge]
- a figurative designation; (Skt. *upacāra*; Tib. *nye bar 'dogs pa*) [s.hodge]
- Language, speech (*vāc*), which is one of the three kinds of permeation of the store consciousness taught in the *Mahāyāna-saṃgraha*. 〔攝大乘論 (T 1593.31.117c3)〕 [cmuller]
- The usage of language to teach the dharma (*deśanā*). [cmuller]
- Language as synergistic with the mental realm of phenomenal differentiation (*abhilāpa*). [cmuller]

[Dictionary References]

Iwanami Bukkyō jiten 239, 293
Bukkyōgo daijiten (Nakamura)429b
Ding Fubao
Buddhist Chinese-Sanskrit Dictionary (Hirakawa)1072
Bukkyō daijiten (Mochizuki)(v.9-10)1043b
Bukkyō daijiten (Oda)582-1
Sanskrit-Tibetan Index for the Yogācārabhūmi-śāstra (Yokoyama and Hirosawa)

One distinctive feature that you will notice in this example, that one does not yet see in standard reference works, is that attribution is not simply given for the entire entry as a unit: responsibility is acknowledged for each segment (XML node) of the entry--and as often as possible, with the equivalent Sanskrit or Tibetan. Both characteristics are especially helpful for those who are doing research and translation. Of course, using XML like this, we can display more detailed information if we want to. But whether or not we decide to display it when we publish the HTML files, the users have the option of viewing the XML source data if they like. For the above-shown entry, the XML source data looks like this:

<entry ID="b8a00-8aaa" added_by="cmuller" add_date="1997-09-15" update="2003-10-11" rad="言" radval="07" radno="149" strokes="00">

<hdwd>言說</hdwd>
<pron_list>
<pron lang="zh" system="py" resp="cmuller">yánshuō</pron>
<pron lang="zh" system="wg" resp="cmuller">yen-shuo</pron>
<pron lang="ko" system="hg" resp="cmuller">언설</pron>
<pron lang="ko" system="mc" resp="cmuller">eonseol</pron>
<pron lang="ko" system="mr" resp="cmuller">ŏnsŏl</pron>
<pron lang="ja" system="kk" resp="cmuller">ゴンゼツ</pron>
<pron lang="ja" system="hb" resp="cmuller">gonzetsu</pron>
</pron_list>
<sense_area>
<trans resp="s.hodge">verbal expression</trans>
<sense resp="s.hodge">(Skt. <term lang="sa">vyavahāra</term>; Tib. <term lang="bo">tha snyad</term>)</sense>
<sense resp="s.hodge">
<trans resp="s.hodge">expresses, recounts</trans>; (Skt. <term lang="sa">abhi-lap*</term>; Tib. <term lang="bo">brjod par 'gyur ba</term>)</sense>
<sense resp="s.hodge">
<trans resp="s.hodge">expressing, an expression, an utterance</trans>; (Skt. <term lang="sa">abhilāpa</term>; Tib. <term lang="bo">brjod pa</term>)</sense>
<sense resp="s.hodge">
<trans resp="s.hodge">discourse</trans>; (Skt. <term lang="sa">kathā</term>; Tib. <term lang="bo">gtam</term>)</sense>
<sense resp="s.hodge">
<trans resp="s.hodge">a figurative designation</trans>; (Skt. <term lang="sa">upacāra</term>; Tib. <term lang="bo">nye bar 'dogs pa</term>)</sense>
<sense resp="cmuller">Language, <trans resp="cmuller"><term lang="en">speech</term></trans> (<term lang="sa">vāc</term>), which is one of the three kinds of permeation of the store consciousness taught in the <title>Mahāyāna-saṃgraha</title>. <bibl type="canoncite"><cit><xref idref="b651d-5927-4e58-8ad6">攝大乘論</xref> <biblScope source="T" div="num.vol.pg-col-line">T 1593.31.117c3</biblScope></cit>
</bibl></sense>
<sense resp="cmuller">The usage of language to teach the dharma (<term lang="sa">deśanā</term>). </sense>
<sense resp="cmuller">Language as synergistic with the mental realm of phenomenal differentiation (<term lang="sa">abhilāpa</term>). </sense>
</sense_area>
<dictref>
<dict><title>Iwanami Bukkyō jiten </title><page>239, 293</page></dict>
<dict><title>Bukkyōgo daijiten (Nakamura)</title><page>429b</page></dict>
<dict><title>Ding Fubao</title><page/></dict>
<dict><title>Buddhist Chinese-Sanskrit Dictionary (Hirakawa)</title><page>1072</page></dict>
<dict><title>Bukkyō daijiten (Mochizuki)</title><page>(v.9-10)1043b</page></dict>
<dict><title>Bukkyō daijiten (Oda)</title><page>582-1</page></dict>
<dict><title>Sanskrit-Tibetan Index for the Yogācārabhūmi-śāstra (Yokoyama and Hirosawa)</title><page/></dict>
</dictref>
</entry>

IV. Making Contributions

An important future enhancement for the DDB will the development of an input form for contributors, to allow them to readily add new entries, or modify presently existent ones. For the time being however, lacking a formal apparatus for the input of new materials, we have been adding material received in attached mail files, mostly in MS-Word format. As long as the contributors use a format with a uniform structure, and are able to submit their materials in Unicode, using Unicode-mapped diacritics and East Asian characters, there is not that much else that needs to be done, as we are able to do much of the main markup with macros and various scripts. We do, however, encourage users to submit their materials with XML markup to whatever extent they are able to handle it, ranging from a minimal type of markup, up to a fully marked-up document using our DTD. On our web site, we offer users the following options (from *http://www.acmuller.net/ddb/notes/Basic_Formatting.html*):

Basic Formatting Suggestions for DDB Entries

Topics:

Updated 2004.09.05

A. Introduction

First and foremost, please understand well: **the usage of XML tagging is not necessary for contributing to the DDB. We will happily accept contributions in popular word processor file formats with no XML markup whatsoever.** If, however, you are interested in going a step or two beyond that, and would like to learn something about how we encode our materials, then please read on.

B. Basic DDB Entry Format

Up to now, the basic organization of a DDB entry has been like this (with some abridgments for the sake of simplicity):

Headword: (Han characters)

Pronunciations:

Chinese (Pinyin):
Chinese (Wade-Giles):
Korean (Hangul):
Korean (Ministry of Education System):
Korean (McCuneReischauer):
Japanese (Katakana):
Japanese (Hepburn):
Translation: (Simple, short-phrase equivalent of the headword, if available)
Explanation: (Detailed explanation of the entry headword)

If you were adding a term, you would type the Chinese next to "Headword." You would then add the pronunciations for the languages you know. Someone else can supply the readings for the languages you can't handle. After the pronunciations, we usually make an attempt to offer one (or up to a few) common renderings of the term. If it were a person, place, temple, etc., we would just supply the commonly used name, such as "Zongmi," "Dongshan," "Jinglingsi," etc. If it were a concept, "middle way," etc. This is followed by a detailed explanation, which can have multiple nodes for multiple contributors, as necessary.

Let's look at example. This is an entry regarding the Korean monk Iryŏn. It is an entry for which I provided minimal information many years ago, and which badly needs to be expanded. But its present brevity makes it useful here:

Headword: 一然

Pronunciations:

Chinese (Pinyin): Yīrán

Chinese (Wade-Giles): I-jan

Korean (Ministry of Education System): Iryeon

Korean (McCuneReischauer): Iryŏn

Japanese (Hepburn): Ichinen

Translation: Iryeon

Explanation: (1206-1289) An important Goryeo monk. A prolific writer, who is most famous for his Samguk Yusa [Chinese title here], a collection of facts and anecdotes which is a basic text for the study of the history of Korean Buddhism.

C. Basic DDB Entry Format: XML

Now, for XML. Rather than starting off with an explanation of XML theory, I think it is simpler if I just re-present the above example using a simplified form of XML.

<entry>

<hdwd>一然</hdwd>

<pron_list>

<pron>Yiran</pron>

<pron>I-jan</pron>

<pron>Iryeon</pron>

<pron>Iryŏn</pron>

<pron>Ichinen</pron>

</pron_list>

<trans>Iryeon</trans>

<sense> (1206-1289) An important Goryeo monk. A prolific writer, who is most famous for his <title>Samguk Yusa</title> 三 國遺事, a collection of facts and anecdotes which is a basic text for the study of the history of Korean Buddhism.</sense>

</entry>

If you look at this for a minute, you will see that there is not much difference between the first example and the XML-tagged example. The basic difference is that here we are using opening and closing tags to delimit information. You will notice that inside the <sense> tags, the title of Iryeon's text, *Samguk Yusa*, is enclosed with the tags <title></title>, indicating that this is the name of written work. We also use similar tags for <term>technical terms</term>, <foreign>foreign words</foreign> and other elements. When this entry is published as HTML, these words will automatically be italicized. We can also use these tags to build indexes.

If can cooperate by using this simple level of XML structuring, it would be greatly appreciated. But once again, it is not absolutely necessary for the task.

D. Basic DDB Entry Format (Fully Developed XML Markup)

The above example shows the barest XML framework—what are called ELEMENT tags. The tags <entry>, <pron>, <title>, etc. are all known as "elements" in XML parlance. But elements can also be enhanced by a very useful secondary layer of information, which is known as ATTRIBUTE information. Please see the same entry, again presented in a manner much closer to the way it is actually contained in our data set:

<entry added_by="cmuller" add_date="1990-09-21" update="">

<hdwd>一然</hdwd>

<pron_list>

<pron lang="zh" system="py" resp="c.wittern">Yīrán</pron>

<pron lang="zh" system="wg" resp="cmuller">I-jan</pron>

<pron lang="ko" system="mc" resp="cmuller">Iryeon</pron>

<pron lang="ko" system="mr" resp="cmuller">Iryŏn</pron>

<pron lang="ja" system="kk" resp="cmuller">イチネン</pron>

<pron lang="ja" system="hb" resp="cmuller">Ichinen</pron>

</pron_list>

<sense_area>

<trans resp="cmuller"><person_entry loc="ko">Iryeon</person_entry> </trans>

<sense resp="cmuller"> (1206-1289) An important Goryeo monk. A prolific writer, who is most famous for his <title lang="ko">Samguk Yusa</title> 三國遺事, a collection of facts and anecdotes which is a basic text for the study of the history of Korean Buddhism.</sense>

</sense_area>

</entry>

I believe that the point of most of the attributes should be obvious, but one of the most important that I would like to draw your attention to is that of "resp", which means "responsibility"—thus, "accreditation." Far distinguished from paper publishing counterparts, the usage of XML in a digital reference work allows us to give credit to the person responsible for every small part of the <entry>. Thus, if someone wanted to add another <sense> element (or "node") to this entry, it could easily be done, giving that person credit in the "resp" attribute.

Also commonly used in the DDB is the "lang" attribute, which tells us the language of the text or foreign word that will be italicized. For texts, we also have a "prov" (provenance) attribute. For temples and geographical entries, we have a "loc" (location) attribute. There are a number of others as well.

Using attributes allows for all kinds of programming possibilities, including various font transformations on presentation, creation of detailed indexes, and so forth.

However, once again, for those for whom this is a headache, it is fine if you want to terminate your exposure to XML here. Ensuing discussions will go into a bit more detail on XML for those who are interested, so you may ignore these if you wish.

V. Near-Term Future Prospects for the DDB

During the past year, we have reached a distinctive new stage in the development of the DDB, wherein suddenly a large number of recognized scholarly experts in Buddhist Studies have begun to contribute data, and major university libraries have decided to subscribe. We presently have some 5,000 entries on the queue awaiting input, with new contacts from interested scholars coming weekly. Thus, we appear to be on the verge of being able to declare the DDB project a major success. We still eventually need to figure out a way to round out the balance of the coverage, so that there is more equal representation in terms of sects and cultural traditions, but this can probably be solved by the attainment of another significant grant or two. But if we can continue to grown at the rate of 5-10,000 terms a year for the next five years or so, it will probably be the right time to begin to turn our full attention to the proper completion of the sister project of the

DDB--the CJKV-E dictionary.