Journal of the International Association of

Buddhist Studies

JOHN R. MCRAE

Zenbase CD1. Computer software produced under the direction of Urs App. Kyoto: International Research Institute for Zen Buddhism, 1995.

There have been a number of projects undertaken in the past few years for the computer input of Buddhist texts, but the most commendable of these is that undertaken by the International Research Institute for Zen Buddhism (hereafter IRIZ), under the leadership of Associate Director Urs App, in Kyoto, Japan. Following the intellectual guidance of Professor Yanagida Seizan, the founder of IRIZ, App has published a very informative journal in English and Japanese, whimsically entitled *The Electronic Bodhidharma*, explaining various issues of computer text input in Zen studies. Now he has published *ZenBase CD1*, which includes the best of *The Electronic Bodhidharma*, a large number of Zen texts in electronic form, plus a handfull of extremely useful tools for electronic text manipulation. The contents of *ZenBase CD1*, plus several updated files, are available at the IRIZ World Wide Web site (http://www.iijnet.or.jp/iriz/irizhtml/irizhome.htm); in addition, this review and certain additional tools and information are available from my web site at Indiana University. (The online version of this review includes all the Internet references as active hyperlinks. See http://www.easc.indiana.edu/jmcrae/ZB/ZBreview.html). Since the IRIZ web site has been updated somewhat since the publication of *ZenBase CD1* (but most recently only in late May 1996, ten months prior to this writing), users are encouraged to explore it and the other sites listed here for more up-to-date information and tools.

Although there are now a number of online sites that provide access to electronic Buddhist texts (see in particular the web site of the Center for Buddhist Studies at National Taiwan University and directed by Prof. Ven. Shih Heng-ching, http://ccbs.ntu.edu.tw/, and the other sites cross-listed there), *Zenbase CD1* is a remarkable and even unprecedented contribution. This is because of the comprehensive and systematic approach taken by App and his colleagues in including not only a major collection of Chan / Zen texts, but also a very significant collection of tools for the manipulation

of those texts in ordinary users' environments, as well as research utilities to assist those engaged in Zen studies. For example, there are tools that allow the reformatting and conversion of text files from Japanese to Chinese encodings; a set of alphanumeric character fonts that allow both Windows and Macintosh users to display and print all the diacritically marked characters used in Buddhist studies; and Christian Wittern's KanjiBase, which provides a systematic method for the specification of characters not found in the most widely used codesets. (The alphanumeric fonts, known as Appeal, should be adopted by Asianists as a convention for the representation of diacritically marked text. KanjiBase is the best interim solution until a modification of Unicode now under consideration is adopted.) The research utilities include a set of bibliographies of English and other writings on Zen studies; a set of Macintosh dictionaries for Zen-related character input and information lookup (presented in a variety of formats for different input methods); and a number of Macintosh dictionaries and Zen lineage diagrams prepared for use with a specific commercial dictionary utility, TSM Passport, for which a demo version is also included. Not all of these tools work flawlessly (even with the updates available on the IRIZ web site), and the documentation provided is frequently insufficient, but their overall value is remarkable.

In addition to these various resources, on which I will comment in specific below, a major reason why *ZenBase CD1* is so important is its underlying design philosophy. Three major principles are involved in this design philosophy: (1) the use of a master / distribution tape metaphor for the preparation of electronic texts; (2) the adoption of conventions for tagging texts developed by and in conjunction with the Text Encoding Initiative (TEI) and the Electronic Buddhist Text Initiative (EBTI); and (3) the use of HyperText Markup Language (HTML) for the simultaneous navigation of both *ZenBase CD1* and the IRIZ web site.

First, let us look at the master / distribution tape metaphor. In articles previously presented in *The Electronic Bodhidharma*, App has explained the metaphor that governs his approach to text archival, which is that of "master" and "distribution" tapes used in the music recording industry. That is, where a commercially available tape of, say, Carl Orff's "Carmina Burana" or the Jerry Garcia Band would contain two sound tracks for reproduction on ordinary home stereo equipment, such a distribution tape would be mixed from a master tape having 17 or 20 tracks or more. Similarly, whereas individual computer users may only have the capacity to read characters found in either the Big5 code (used for traditional characters in Taiwan and elsewhere) or the JIS code (used in Japan), this will probably

not be sufficient to represent all of the characters used in a given Zen or other Chinese Buddhist text from centuries past. Therefore, the producer of the master text should take pains to include as much information as possible about the original printed text, including the precise specification of character variants, so that as users' technical capabilities improve (as they inevitably will, in part through the efforts of IRIZ and other such institutions) they can use more and more accurate distribution texts. In addition, the master version of any electronic text, rather than the less accurate but certainly usable distribution versions, should be that used for the creation of more complex and sophisticated database tools, such as are certainly going to appear in the near future.

Second, *ZenBase CD1* is the first Buddhist input project to adopt the conventions of tagging developed in conjunction with the Text Encoding Initiative (TEI) and the Electronic Buddhist Text Initiative (EBTI). Simply put, these conventions involve the inclusion in each electronic text of (a) a header explaining the source of the text, the people responsible for its creation, the standards of accuracy achieved, the specification of the codeset and additional characters used, and definition of tags used, etc. (part of this header is known in TEI terminology as a Document Type Definition [DTD]); and (b) tagging of specific items within the text using Standard Generalized Markup Language (SGML) in ways conventionalized by TEI and EBTI. Since there are obvious advantages to having information about the provenance of any electronic text attached to the text itself, the header may be included even if there is no SGML tagging applied. In keeping with this rule, every electronic text provided in *ZenBase CD1* includes such a header, but due to the fact that the conventions of TEI / EBTI tagging for Buddhist texts are still in a very early developmental phase, only one text on the CD has such tagging. In fact, there are three categories of texts included on the CD, one that includes a first attempt at TEI / EBTI tagging (represented by only one very substantial text), one that includes texts which have been thoroughly proofread and which are believed to be extremely accurate (represented by some 16 texts), and a third that includes texts which have been input and proofread only to a level of moderate reliability (represented by some 58 texts).

Third, the explanatory and instructional files on both *ZenBase CD1* and the IRIZ Web site are all written using HTML, so that they may be navigated using any of the many Web browsers available today, such as Netscape Navigator, Mosaic, or (for those without graphical interfaces) Lynx. In fact, all these files are available in two versions, one in English and the other in Japanese; it would certainly be appreciated if IRIZ would

make available Chinese-language versions of these same files at some point in the future. (I have prepared a Chinese translation of App's article on preparing electronic editions; see http://www.easc.indiana.edu/~jmcrae/ZB /App_msinput.Big5.html.) Using recent versions of Netscape Navigator or Mosaic, the user needs simply to have Japanese capabilities available on his computer and to select the appropriate Japanese font. (In Netscape 3.0 for the Mac, for example, this is done by pulling down the "Options" menu, selecting "General Preferences" and then "Fonts," and choosing a suitable Japanese font. Other Web browsers will have procedures that are conceptually similar.)

   ZenBase CD1 even includes both Netscape 1.1N for the Macintosh and Mosaic for Windows. These versions are now long since out of date, and one suggestion is in order for users wanting to install Mosaic for Windows from the CD: my attempts to follow the installation procedure on three separate PC clones (two no-names and a Compaq) were unsuccessful until I realized that I had to copy the files from *ZenBase CD1* to a temporary file on the PC and run the installation procedure from there. In any case, those with Internet access should download a more recent version of Netscape Navigator via ftp from ftp.netscape.com, and those without Internet access should be able to find a version of either Netscape or Mosaic (for either PC, Macintosh, or Unix machines) at a computer store or bookstore, or through a mail order house. In any case, the problems just mentioned are indicative of the one shortcoming of *ZenBase CD1*, which is a lack of accurate, meaningful, and/or sufficient instructions for many of the tools provided.

   There are three categories of tools found on *ZenBase CD1* : (1) font utilities, especially the Appeal font, for diacritically marked characters, and KanjiBase, which provides a means by which to handle characters not in the Big5 character set; (2) tools specifically relevant to electronic text creation, most of which utilize PERL or MacPERL; and (3) other useful tools not necessarily directly related to electronic texts per se.

1. I am placing Appeal and KanjiBase together because, although quite different from each other, they both resolve difficulties involving fonts. Appeal is a font (represented in both Mac and Windows versions) that should be used as a standard convention for alphanumeric text by everyone in Buddhist studies. It is based originally on the fonts created by K. R. Norman for use by Indologists, but it has been substantially improved by people at the Hōbōgirin (where it has been used for the preparation of recent editions of the *Cahiers d'Extrême-Orient*) and IRIZ. Those who ac-

cess my Web site will find that I follow my own advice: users wishing to see diacritically marked text displayed correctly are instructed to download and install Appeal on their systems. The differences between Appeal and the Norman fonts are too technical to be discussed here; see http://www. easc.indiana.edu/~jmcrae/fonts/Norman+Appeal.html for a discussion and references. Although font loyalties tend to approach the strength of religious conviction, for Buddhist studies Appeal has the following advantages: it works better with a greater number of Asian and European languages, it displays better in both roman and italic styles, and it is technically better suited to work on a variety of platforms. Fortunately, for the great majority of users who have to convert files from one font to the other, only one annoyance occurs: the need to do a search-and-replace for ū (long u, lower case).

The KanjiBase system, developed by Christian Wittern, is currently available only for Windows users, although I have used it quite successfully on the Mac under SoftWindows. Although it will probably be superseded in a year or so by a modification to the Unicode standard recently proposed by Prof. Hsieh ChingChun of the Institute of Information Science at the Academia Sinica in Taiwan, this is an important innovation that should be adopted as at least an interim solution by every project for the input of Chinese texts, Buddhist or non-Buddhist. What KanjiBase does is provide a sensible and reliable manner for the treatment of characters not found in the Big5 character set. Now, Big5 includes about 13,000 characters, which is generally adequate for ordinary written Chinese communication but not sufficient for the representation of classical Chinese texts. Although the number of characters in any given text that do not occur in the Big5 code is small (probably no more than 3-5% of the total character representation of any text, and often less than 1%, according to statistics compiled at IRIZ), the requirements of accuracy demand that these characters be represented in some fashion.

Most Chinese software systems allow the user to create and specify a small number of extra characters, referred to as *waizi* in Chinese or *gaiji* in Japanese. However, since each text input project has defined its own set of *waizi / gaiji*, it is extremely difficult if not impossible to exchange data between different computers. In theory, conversion tables could be constructed, but they would certainly be "lossy" (since sets of extra characters created by different users would never correspond to each other exactly), and in practice this is hardly ever done. Also in theory, it might be possible to use an entirely different code from Big5 as a master code, such as the CCCII code used by libraries in Taiwan, the closely related REACC code

used by the U.S. Library of Congress and comprehensive library databases in the U.S., or the recently developed but still underused Unicode. For different technical reasons explained by Wittern in an article printed in *The Electronic Bodhidharma* and included in *ZenBase CD1* (see *The Electronic Bodhidharma*, no. 4 (1995), or the file CEFINTRO.HTM on the CD or IRIZ web site), none of these codes is acceptable.

KanjiBase resolves this problem by the combined use of the Big5 code (actually, the most commonly used variant of Big5, that adopted by the ETen operating system) and the CNS (for Chinese National Standard) code. Although CNS is officially promulgated by the government of the Republic of China, it has never been adopted by any Chinese software or system creator. However, it has the advantage of specifying a codeset of some 48,000 characters in a fairly rational fashion, with relatively little overlap and omission. Wittern has designated ETen Big5 as the first "plane" of his system, with five CNS planes and three additional planes (to allow for future growth) residing on top of this. (As a computer standard, CNS involves features other than codepoint definitions which ae irrelevant.) In order to find a character variant, one enters either the radical number and stroke count or a four-corner code, then uses arrows to select the specific character desired. When a character is selected, the program returns a larger bitmapped image and other information, which includes some but not necessarily all of the following: Chinese pronunciation in Pinyin, Japanese pronunciation, English meaning, and numbers in Morohashi's *Dai Kan-Wa jiten*, Big5, Unicode, and CCCII. Macros to allow the bitmapped image to be introduced into a Microsoft Word 6.0 file were promised on *Zenbase CD1*, but are actually only available online from Wittern's own web site. Also available there are updated Big5<->SJis conversion utilities and an improved Windows version of the Appeal font; for all of these see http://www.gwdg.de/%7Ecwitter/bugfix.htm. Most promising, however, is that this site features an online implementation of KanjiBase, although unfortunately at this point it is still of limited availability. (I have not yet been able to access it.)

It is unfortunate that KanjiBase is not available yet for the Macintosh platform, but it is possible to use it on a Mac under SoftWindows. I have successfully run KanjiBase under SoftWindows on a PowerBook 280c, but there are a couple of provisos to be made. First, the program should be copied from the CD while operating under SoftWindows; I first copied it to a directory accessible by SoftWindows using the Macintosh Finder, and it just didn't see some of the most important files. (It seems that they are experienced by the MacOS as hidden files, a technical oddity of *ZenBase*

CD1.) Second, make sure to copy the program to a hard disk file named D:, since it just won't run from disk drive C: This last piece of advice should apply to those running under native Windows environments as well.

As this review goes to press I have just learned of the proposal for a modification to Unicode by Prof. Hsieh, already mentioned above. The recommendation is that characters not in the Uni-Han area of Unicode (that reserved for CJK characters) should be represented by a combination of codes representing the individual elements used to write the characters and combined in the order the character is written. In this fashion Unicode would be spared the necessity of designating specific codes for the tens of thousands of characters (estimated in the 160,000-200,000 range) that are used only very rarely (less than 0.1% of the time, on average), but at the same time a unique code could be generated for each such character shape. This proposal is now under consideration by the Unicode Consortium, but until it is adopted KanjiBase can provide a very workable (and eventually easily upgradeable) solution. And, if KanjiBase is in fact superseded, one-for-one correspondences between the characters it covers will allow for that Valhalla of computer technology, an easy upgrade path.

Also included in *ZenBase CD1* is an implementation of Werner Lemberg's CJK TeX, a platform-independent implementation that allows the typesetting of documents that can include Chinese, Korean and Japanese text. This implementation of CJK TeX includes support for CNS via the KanjiBase code references. This may be wonderful news to some people, but I'm a FrameMaker kind of person: I have never taken the time to learn how to use TeX, or even LaTeX, which are dearly beloved of true computerphiles but just too far from WYSIWYG (what-you-see-is-what-you-get) for me. (Just as, prior to moving to Cambridge I had a nightmare in which I actually *liked* the Celtics, I dread the day someone convinces me to take the plunge into TeX). Hence, I cannot comment on this system.

2. The tools most closely related to electronic text creation found on *Zen-Base CD1* include a number that utilize PERL, and there are instructions for installing PERL (or MacPERL) on either a PC or Macintosh computer. (Mac users should note that the "pyramid" shaped icon referred to in the instructions does not occur in the CD versions of the Mac environment, where it appears—for reasons only the original program authors under-stand, no doubt—as a camel on wheels. This has been changed in the updated tools to include both camel and pyramid.) These PERL-based tools include the following:

A. A concordance maker, which "creates a complete concordance from a text file (JIS or Big-5)" which in conjunction with related Word 6 macros can be used to print "a perfectly formatted and paged concordance." There were substantial problems in this tool as published on the CD, and a newer version has been installed on the IRIZ web site. I have not tested this utility.

B. Kanji (Chinese character) code conversion tools, which can be used to convert text files from Big5 to JIS or vice-versa. Code conversion is a very inexact process due to the differences between codes (i. e., one code may have nothing to correspond to a certain character in another code, or it may have more than one character). As a result, *ZenBase CD1* provides different tools, with three degrees of strictness of conversion.

Curiously, though, *ZenBase CD1* provides tools only for conversion between Big5 and JIS codes. Conversion utilities that handle additional codes (most significantly, the Guo Biao or GB code used in the People's Republic of China) are easily available and could easily have been included. The interested PC or Mac user can use ftp or Fetch, etc., to access ftp.ifcss.org:/software/ or its mirror site ftp.cuhk.hk:/pub/chinese/ifcss/software (or use a web browser to get to http://www.ifcss.org/software/) for Sinocode.zip (for the PC) or Hanzi Converter (HC 1.5 or HC 3.0, for the Mac; I actually find HC 1.5 to be more convenient). Also check Carlos McEvilly's Chinese Language Homepage (http://www.netcom.com/bamboo/chinese) for his Bamboo Helper and a great deal more. I have worked with both Sinocode and Hanzi Converter to convert files written using the WPS program so widely used in the PRC to Big5 files; both programs need the source file to be cleaned of extraneous format codes. Recently, I've found that for my purposes the import capabilities of the shareware text editor Unicorn (available through ifcss.org) are more convenient than any of these stand-alone GB <->Big5 conversion utilities, and Unicorn is convenient in other ways as well.

C. Kanji normalization tool (Big5 texts only), which "normalizes a text file for more convenient use in ordinary text search or in concordances; the program also generates a record of all conversions that were effected." I have not used this tool, but the problem it addresses is a real one. Since the Big5 code contains variant glyphs of some characters at different codepoints, this program should convert the variants into the codepoints most commonly used.

D. Text format conversion tool, which "converts the format of text files among the RAW, APP, and TAB formats." I have not tested this tool; the

three formats are those used for different purposes within the IRIZ text archives.

E. IRIZ text statistics tools, which can be used "to count characters and compounds and to generate gaiji information lists."

F. The search utility IRIZGREP, for performing text searches within a file or set of files, for which a new version is offered on the IRIZ web site. This works very nicely indeed, and is perhaps the tool users will find most useful.

3. There are a number of tools included on *ZenBase CD1* which, although not directly related to electronic text creation, are nevertheless very useful. These include:

A. Versions of Netscape 1.1N and Mosaic 2.0 for the Mac. See the discussion above regarding installation of these or more up-to-date versions.

B. Three Mac system utilities, FontPatchin', SearchFiles 1.3, and UltraFind 1.0.3. There is no explanation of how the people at IRIZ find FontPatchin' useful (documentation, or at least a brief statement, would have been helpful), but Charles Wivell, of the University of Rochester (retired), has told me that it is useful for including Chinese and Japanese characters in electronic mail. However, Wivell warns that the FontPatchin' control panel causes some problems with ordinary computer operations, so that he disables it by putting it into a "Control Panels (disabled)" folder when it is not needed. SearchFiles (this is the most recent version) allows one to search all files in a given directory for the presence of specified text strings. UltraFind is described as allowing one to find specific files or groups of related files anywhere on one's computer or network and move or perform other operations on them; version 2.2 is now available (www.ultradesign.com/ultrafind). Both programs are shareware. In my own experience, SearchFiles has worked nicely (even finding Chinese text, although without appropriate character display), whereas UltraFind has not. That is, given the same search task, SearchFiles returned results and UltraFind did not.

C. BBEdit Lite 2.3.1, a Mac editor intended for use in writing HTML text (version 3.5.1 is now available on the Internet, and ver. 4.x is available by purchase), and SedMac 1.0. I have not used these.

D. A demo version of TSM Passport, a Japanese dictionary utility for the Mac. After toying with the demo (which is almost thoroughly crippled), I purchased the complete program at a store in the Akihabara electronics

mart section of Tokyo. *ZenBase CD1* provides dictionaries for use with TSM Passport that allow the user to access an extensive bodies of information about Chinese and Japanese Zen. Although I have had some difficulties with this program (I have almost got it running correctly on a PowerBook 280c, but it crashes when used on the PowerMac at home), but if you can take the time to get it to work it is highly recommended. If you don't have someone in Japan to purchase a copy for you at an Akihabara discount price, you can order the program from the developer: Mercury Software Japan, Hassei Bldg., 2F; 20 Rengezō-chō, Shōgoin; Sakyō-ku, Kyoto; Japan 606; (075) 751-0205; fax: 751-0206. The program is listed on the company's web site, but with no information about the current version number, and no order or demo download options. (See http://www.iijnet.or.jp/mercury/WORPRO.html.)