Digital Buddhist Texts and Buddhist Universities[1]

Lewis R. Lancaster[2]

It is a great pleasure to be part of this meeting of the International Association of Buddhist Universities. I appreciate the invitation to speak today about the role of IABU and the work that has been done with regard to digital Buddhist texts. Bangkok is an appropriate place to discuss these issues because on 30 May, 1988, The Digital *Tipitaka* Development Team at Mahidol University Computing Center, Thailand announced the completion of the first major project to digitize Buddhist texts. The Siam edition of the Pali *Tipitaka* in forty-five volumes had been successfully digitized and released. It is now 20 years later and since that time new projects have come into existence including the Pali version of the *Chattha Sangayana* edition in Roman and Burmese script, the *Buddha Jayanti Tiptaka* of Sri Lanka, the Pali Text Society Edition of U.K., the Tibetan edition by the Asian Classics Input Project, Buddhist Sanskrit Text Project of University of the West and the Nagarjuna Institute

[1] Presented at the 1st IABU Conference on Buddhism and Ethics at Mahachulalongkornrajavidyalaya University Main Campus, Wang Noi, Ayutthaya, Thailand in September, 2008.

[2] Lewis Lancaster, PhD. (Wisconsin), was professor at the University of California, Berkeley & president of University of the West, Rosemead, CA, USA; he is also the founder and Director of the Electronic Cultural Atlas Initiative (ECAI). (www.ecai.org). His works on East Asian Buddhism and digitization of Buddhist canonical texts.

of Nepal, Koryo Edition of the Chinese canon known as *Tripitaka Koreana*, and the *Taisho Issaikyo* edition available from CBETA at the Dharma Drum Buddhist College in Taiwan and SAT in Tokyo University. The digital versions of the Chinese, Sanskrit and Tibetan canonic material have altered the landscape of scholarship in the field of Buddhist Studies. This immense effort to produce the databases for the texts has also created a challenge and an opportunity for Buddhist universities.

If the digital age is to fulfill its promise, it requires a well-coordinated effort to avoid incompatible platforms, codes, categorization systems and unnecessarily repeated work. Over the past two decades, there has been no organization such as the International Association of Buddhist Universities, to help with this effort. The work of input and the creation of digital Buddhist materials has been done by institutions and individuals who have often worked alone and with only local funding. There are a host of indispensable issues that need to be addressed by educators within the Buddhist communities. One of the most pressing is the development of digital research and reference tools for the large datasets.

With the growing number of online resources, the Internet has become the first choice of many students and scholars. It is crucial that these online materials provide the best and most accurate information. This shift from paper to digital resources is already an established reality and concerted action in providing leadership for the creation and appraisal of this material will determine the success of the new medium. This moment in history calls for prompt and definite action from an organization such as the IABU.

There is a danger that we will become complacent about the digital Buddhist text material, feeling that the input has been completed and there is no need for further work. We must be aware that digital data is the most fragile format for information ever invented. It can disappear in an instant and beyond recovery. I think of digital data sets as being like a baby that never grows up, never moves beyond the need for support, never moves to a new location without extreme effort on the part of the creator, and is always susceptible to viruses. We have these wonderful versions of the Buddhist texts in the computer and now we must think and plan

about maintaining them so they will be sustainable for the future. Librarians are the best candidates to do this archiving and preservation and Buddhist universities must take the lead in providing for this necessary effort if our data is to survive. Sustainability is dependent on the coding, software, and formatting of information. Every day thousands of pages of information disappear from the World Wide Web because a server is closed, a project has completed the funded period, individuals retire, campuses stop supporting older software, and no allowance has been made to move the data to new platforms. Who will make certain that all of the Buddhist text input is carefully placed in an archive that will assure it a long life into the foreseeable future? I believe this is a task that IABU must consider carefully.

Scholarship changes with the availability of digital information. Digital library initiatives around the world are providing an amount of data on the web that surpasses what most campuses have in printed books on library shelves. While the data is available, digital libraries have not yet created the tools for the referencing procedures. Our codex libraries have Reference Rooms with librarians who help the users find resources contained in those books. At this time, there is nothing comparable to a Reference Room in the internet environment and it shows in the problems that students have in assessing the value of the information found with a Google or Yahoo search. In order to give the best help possible, it will not be enough simply to point to acceptable sources.

Researchers will need to be given support in understanding the context of the information. This context implies answers to a number of generic questions such as "Where was it done?" "Who did it?" "When did they do it?" For librarians it means that digital material cannot only be indexed as an "object" as our books are catalogued in the codex library. A new kind of cataloging must emerge that marks up every "object" but considers that "object" as an "event." Consider a Buddhist text in digital form, it is important to know it by title and author or translator when appropriate. However, the text seen as an "event" must be marked up to show us the history of what we see on the screen. Which edition do we see, when was it made, where was it made, what version of the edition was used by

those who did the input, where was the input done, who was responsible, what software and procedures were used for the work, etc. The reference catalog for the digital canons will look very different from those of the former card catalogs. Because Buddhist material has its own context, it is going to be part of the task of librarians and scholars within the discipline of Buddhist Studies to provide this background.

As we have easy access to the input projects for the Buddhist canonic texts, users will want tools to aid them in managing the search results, finding images of the original manuscripts or prints where available, cross links between different language versions for the same text, built in dictionaries, and analytic software to determine patterns within the texts.

Our users will want to have more than individual sets of data that must be searched independent of all other sets of data. The word "silo" has been selected to describe the amassing of specific information in one site. In the future, we must have search engines that will find results across many "silos" without the user having to enter and exit each one individually. For Buddhist Studies, this will mean being able to search for a term in all of the canon databases at once and retrieving the information in a variety of forms.

I have been researching this problem for some time and have now developed a prototype interface. It is a prototype only and still has much work that needs to be done and I hope to find others who wish to participate in the effort. The prototype shows us a word or phrase search. (See Appendix I) When the results are returned, they are shown as an image where red dots appear representing the target word within a sea of blue dots that represent the text itself. Images allow us to see either the whole of the canon in one view or details limited by our interests. This immediate display of the patterning across the whole of the canon gives us an idea of whether our word is widely used, rarely used or only used within certain texts. The interface exhibits a window where we can go into the image and ask for the natural language text and we can then show a scanned image of the printed source or manuscripts. The pages of the canon images are shown to us so we can move back and forth between examples. (See Appendix II) If we add analytic software, we can

also see patterns displayed for us in a variety of ways such as one that shows the number of target words according to the time of translation of texts or structural patterns such as Ring Composition. Our context builder for the Chinese version can present the canon by time of translation, order of catalogs, translators, or place of translation. This is just one way to see context building. (See Appendix III)

Researchers will want to have a window in the interface that will allow the appearance of multiple versions of the same section. For example, with the Chinese canon we would like to have a scanned image for the *Taisho Issaikyo* printed page, the Koryo print from the blocks of Hae-in Monastery, manuscript images from Dunhuang, rock cut rubbings from Fang Shan in China, as well as the corresponding passage in Sanskrit, Tibetan, or Pali. For each version, the user may want to draw into the window images of prints, manuscripts, and fragments for any passage.

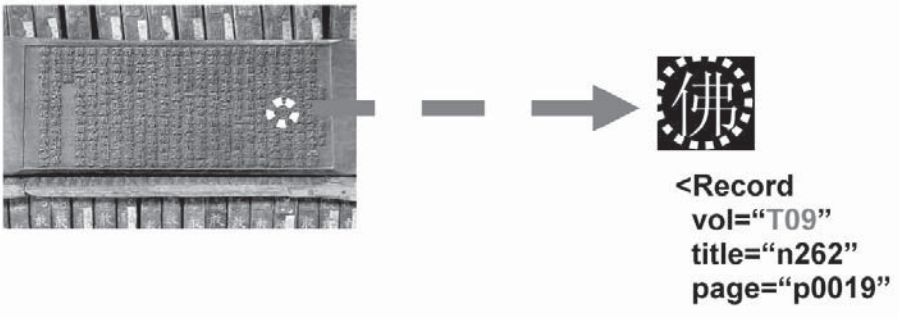
The capabilities described above will change the way in which we edit and translate texts. No longer will multiple footnotes at the bottom of the page indicating alternate readings be acceptable. With the computer, we will ask instead to see an image of the "witness" that shows a different reading. The user will be able to judge for themselves the variety of readings in a way that is much more complete and accurate than relying on notations from one scholar who had access to a number of resources. It has been nearly impossible for readers to challenge the footnotes of editions because there has not been easy or even possible access to the original documents used by the editor. With the new expanded interfaces, we will all have the opportunity to view for ourselves the images of these "witnesses" and for the first time have the ability to "falsify."

I look forward to the future and to the new insights which we can gain by using the computer to help us with pattern identifications that we have never noted before. The comparisons of canonic versions, witness imagery from both prints and manuscripts, and use of databases available to all other scholars opens us a new horizon. The International Association of Buddhist Universities will be an important part of these developments.

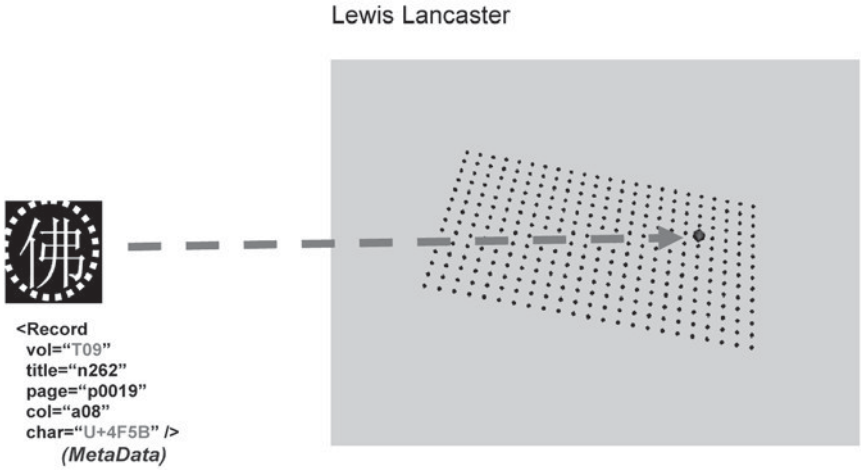
Appendix I

Converting the glyphs that are displayed from the fonts into colored forms.

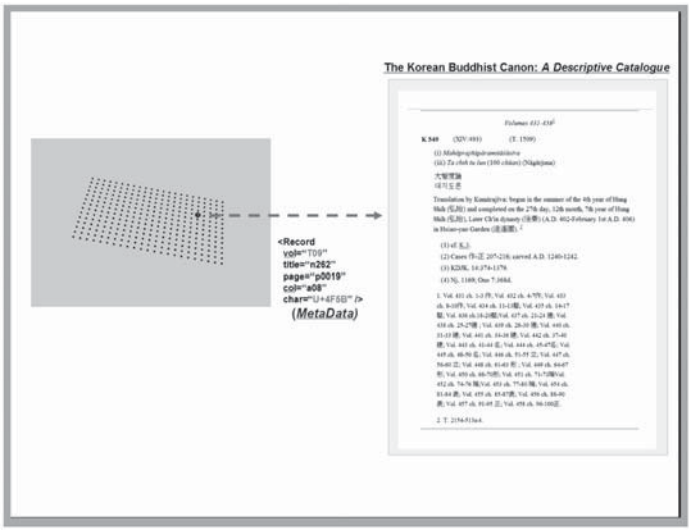
Here we see a "page" of the material, in this case the 13th century printing block, and on it are 23 lines containing 14 characters. All of the material appearing on the more than 83,000 blocks has been digitized in full text. During the process of doing this work, each of the characters was marked up with a Unicode designation such as :



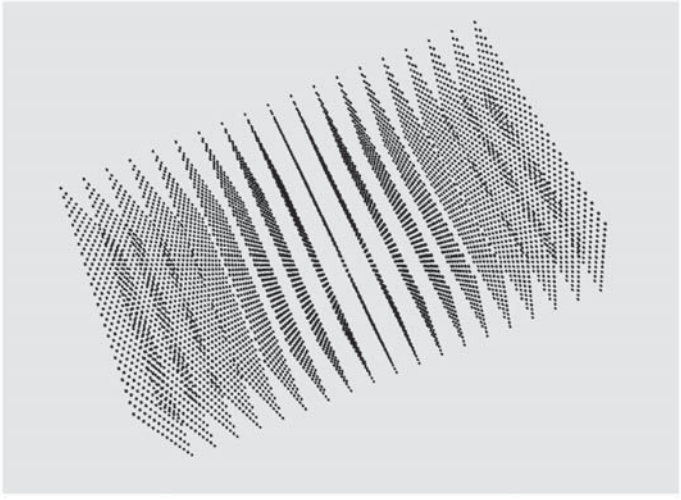
In the proposed project, these millions of glyphs will be converted into an image (a blue dot in this case) that will have the same metadata as the glyph.



Because we have a digital catalogue of the set of texts, it is possible to create a link from each blue dot to the appropriate description of the text in which it appears. In this way, we keep the VR closely tied to reference works and context.



The blue dots representing the glyphs are then arranged in the order of appearance on "pages." Each panel represents a page and each dot a glyph in its proper order.

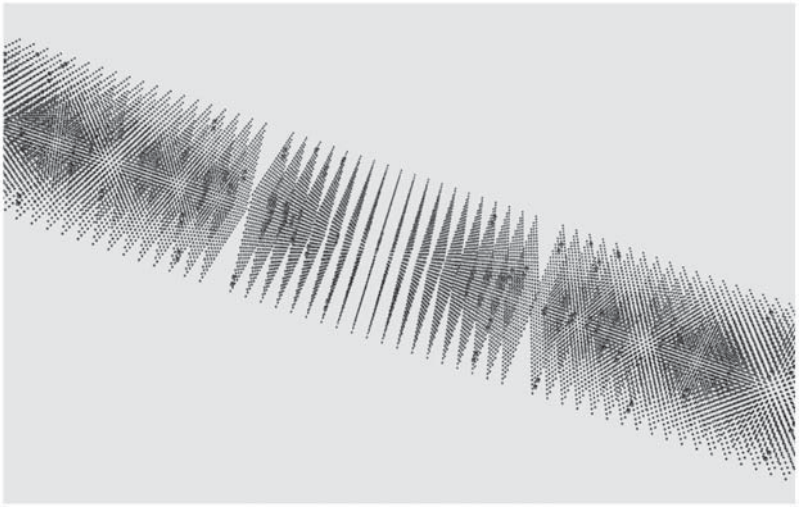


In this way the 3-D and VR software allows us to present these abstracted images of the canon in a recognizable arrangement for the user. At first, they can see “pages” and “columns” is the exact format of the print version. Since the original printing blocks are still housed in a library-like environment, VR can give an added dimension of allowing the user in an immersive environment to “walk” down the “aisles” of the Hae-in Monastery archives in Korea. When this is converted into the VR scheme of this project, users can move from the picture of the blocks directly to the abstracted blue dots representing the words contained on the surfaces.

Searching for terms in the VR presentation

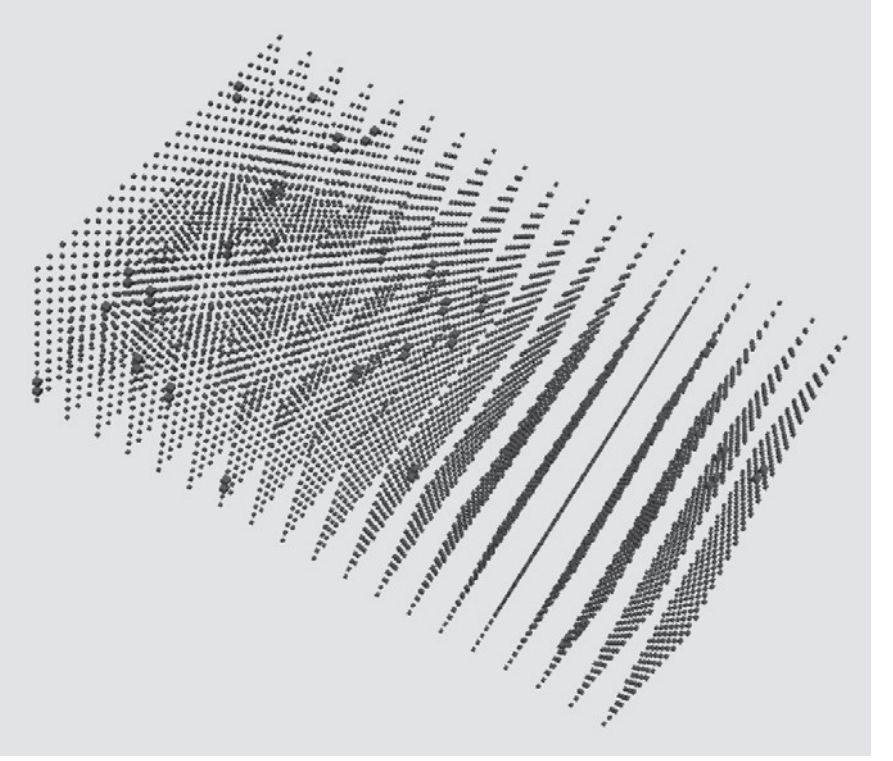
The use of the blue dots for the search for patterns can start with a string search of the digital data. The result of such a search in VR will differ greatly from what we currently see as depicted below.

When a string search is made in the VR medium, the results are seen visually as “signals” on the “pages” as the blue dots of the target word are changed in color and size to indicate the presence of the target word.

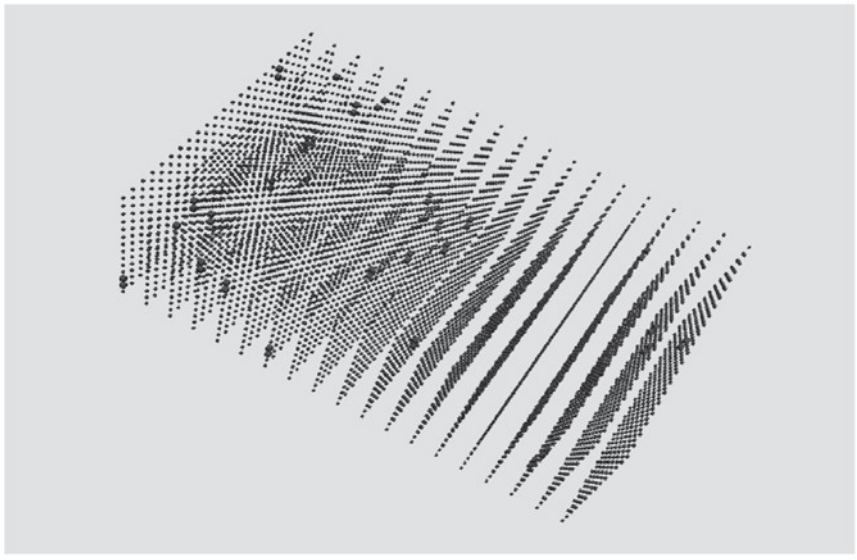


Identifying patterns in search results.

At this point, the user can "see" all of the places where the target word is found. Rather than moving directly to the glyphs that make up the printed version, we remain in the abstract arena for further exploration of patterns that can be presented in images. A first pattern can be word clustering. The user can spot places where clusters are visually apparent.



When we compare the Google menu resulting from a search with the VR dots imaging, certain patterns are easily seen in the latter. It can take many hours for scholars to use standard search results of lines of text references as data for identifying clusters.

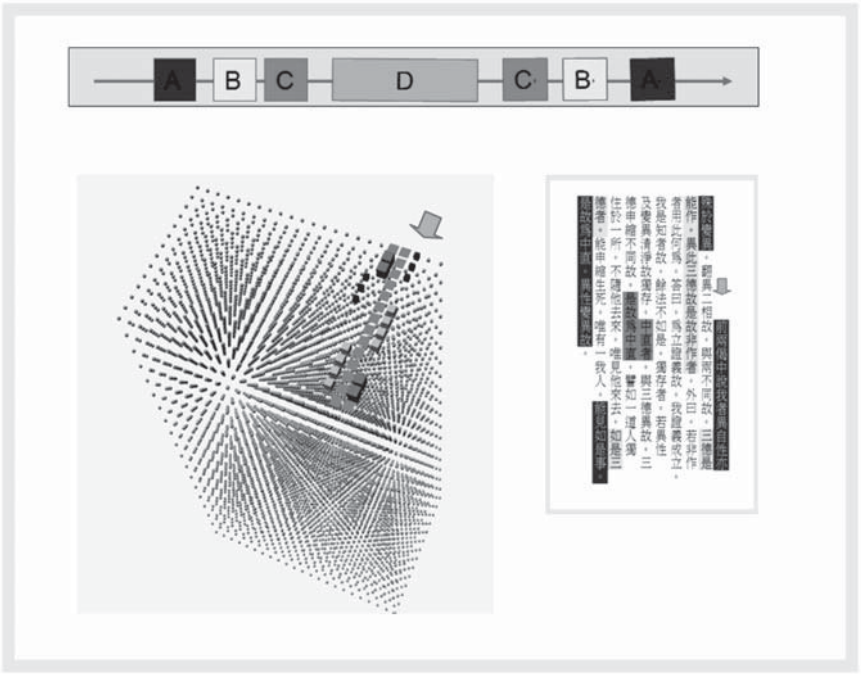


Appendix II

Chiasms and Ring Composition

From Biblical and Classical studies, we are aware that ancient literature exhibits chiastic structure, a repetition of the words in an inverted ordering. In this type of structure, the end corresponds to the beginning. It can be seen as A B C C' B' A'. From this we have "ring" constructions where the story starts, proceeds to a turning point, and then repeats the elements of the story in reverse order until one arrives at the end which is a corresponding or similar statement to the beginning. A recent volume by Mary Douglas Thinking in Circles: An Essay on Ring Composition points out the need to make cross-cultural studies of the particular literary phenomenon of chiastic structuring.

Using the search capacity of the digital version of the Buddhist canon, we can begin to create a VR image that includes the text locations where this ring structure appears. In the illustration below, we see the structure of the Ring Composition with three elements ABC that appear in the first segment and then repeat as CBA in the second. Located between the final element of the first segment and the initial element of the second segment C and C' we find the "kernel" or theme of the ring located at the turning point. When searching for a Ring Composition in the VR abstract format, we see the varied color of the dots in the Chinese reading order of lines from right to left and order of the individual glyphs from upper to lower. The software must search to find places where a phrase is repeated as in the blue dots of A and A'. The second element and third element B and C are discovered by looking for a serial duplication of phrases. The key to the ring structure is that the duplication must appear in reverse order. We see that this is the case in the Chinese text Blue dots (A) followed by Green dots (B) followed by Pink dots (C) with the turning point and "kernel" Orange dots (D). The order after the turning is Pink dots (C') followed by Green dots (B') and completed by Blue dots (C'). From the research on Ring Composition, we should pay close attention to the Orange dots (D) because they represent the theme or major concern of the structure.

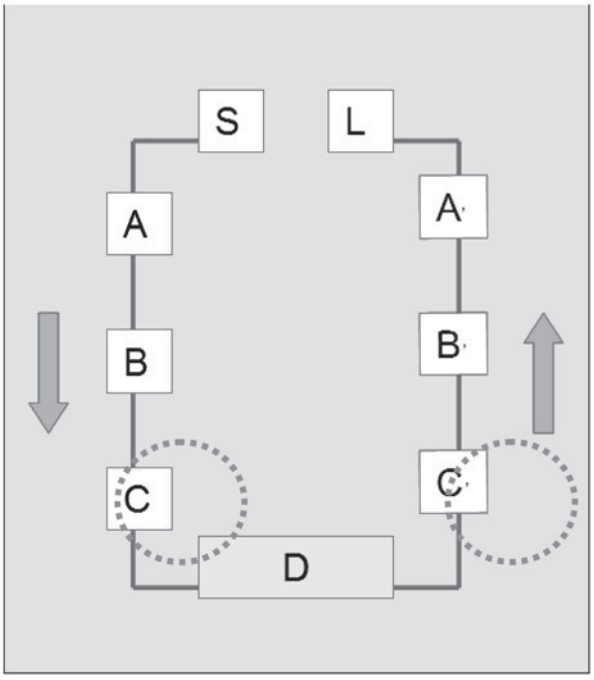


The significance of understanding this ring construction is now seen as crucial for our understanding of literary texts. Without seeing the rings, most ancient material can be misunderstood as chaotic collections of repetitive phrases. From the patterns found by such structure, we can also find a structure made up of a micro-structure composed of numerous minor rings.

Interpreting Ring Composition

Since the Ring Composition is composed of a series of words, concepts, minor rings that appear in a sequence which "turns" and all of the series is repeated backwards, we can look for target words in terms of placement within the structure. If a concept appears as one of the thematic words for the parallels of the ring that gives us an opportunity to deal with the concept in terms of its relationship to the other thematic words. That is, if the target word is B in the sequence of A B C C' B' A' within an identified ring, we can expect the word to have a special relationship to A and C. In order to see the Ring Composition in a clearer fashion, we have the opening statement as S and the repeat of it at the end as L (known

as the "Latch" phrase, or the one that binds the two ends of the ring). S and L are identical statements with different functions of starting and ending the ring. The two encircled elements, C and C' are the bounding that shows the turning of the ring and the delineation of the "kernel" element. We are able to identify the turn because C is the first element to be repeated.



The ring structure has a beginning, a turning point, and an end. We can identify the turning point as the place between the last item of the first sequence and its repetition in the second sequence. That is we have A B C (turning point) C' B' A'. The turning point is the place where the general message of the ring is placed. In the space between the first repetition, we expect to find the "kernel" of the ring.

In the search for a target word, it would be of great interest to see if that word appears as the "kernel" (D) of a ring. That would give it greater weight than if it is just one of the thematic parallels (A B C). Below, we see the imagery where the analysis of the word placement shows us a possible Ring Composition because we find the space between a repeated phrase. Once this space has been located, it is necessary for the software to seek for the structure that exists within this boundary.

Appendix III

The figure below shows he draft interface layout. This interface integrates the various components of the data visualization. By looking at the sections of this interface, we document the progress in data collection and analysis.

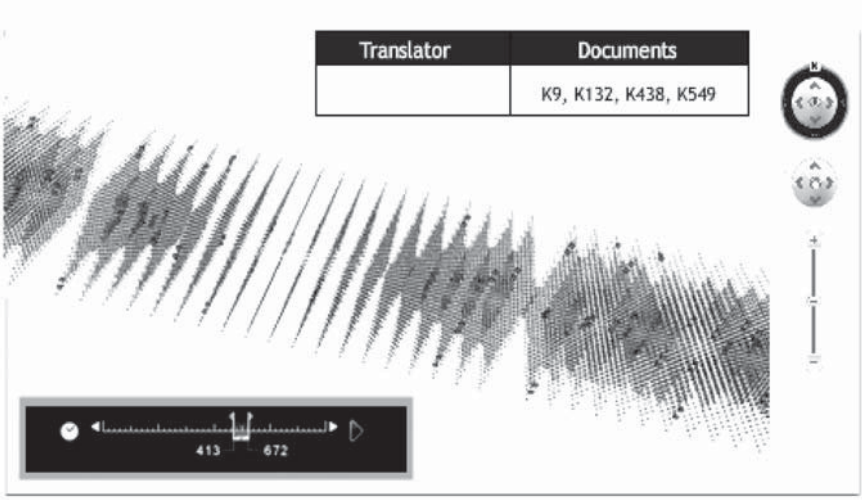
Our first task was to Identify the target source for the data to be used and negotiate the full use of the Koryo Canon in both its ditgital full text as well as the scanned images of the rubbings of the more than 83,000 printing blocks. Having secured this raw data, we reformatted it into a repository suitable for our search engine. The text data comprises 100 MBs and the scanned images of the rubbings from the original printing blocks is 25 GBs in low resolution. Following the acquisition of the canonic material, we constructed out search engine. Based on the Suffix Array Technology, the search engine leverages our implementation from the collaborating group known as CBETA in Taiwan. The strategy in constructing the engine was based on the requirement to be able to make an exhaustive search for any string within the data. Once the repository of data was in place we developed the search engine activated from the first window of the interface.

The first function of the search engine brings up every occurrence with line, count and text title. This window can be viewed by the user but will not be automatically displayed in the interface.

In the illustration below, we show the display offered in the interface to the user.

We have converted Chinese glyphs from natural language form to abstracted "blue dots" each carrying the same metadata as the original glyph. In order to help the user, we arrange these "blue dots" in "pages." We have completed the task of creating a set of 52 million "blue dots" that represent the arrangement of the original Chinese characters for the Koryo printing of the Buddhist canon. These dots are presented in the form of "pages" which can be manipulated by the user.

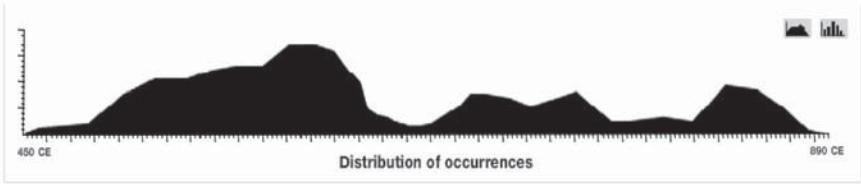
The second level of linkage from the full text search to "Blue Dot" surrogates allows the depiction of the appearance of the target word as a "Red Dot" among the background of all the other characters shown as blue. The linkage is based on the structural metadata of the original text data in terms of text, page, line, position in line, and coding of glyph located in that position. For viewing of the dots, there is the button to the right for 3-D navigation control.



These displays can exhibit complex methods of searching:

Multiple string search (allowing for cluster identification) "String-Pair-in-Distance" (allowing for search where a second occurrence of the target may be distant by many lines in the text). Signature String (displays signal occurrences that may be related to such structure as Ring Composition)

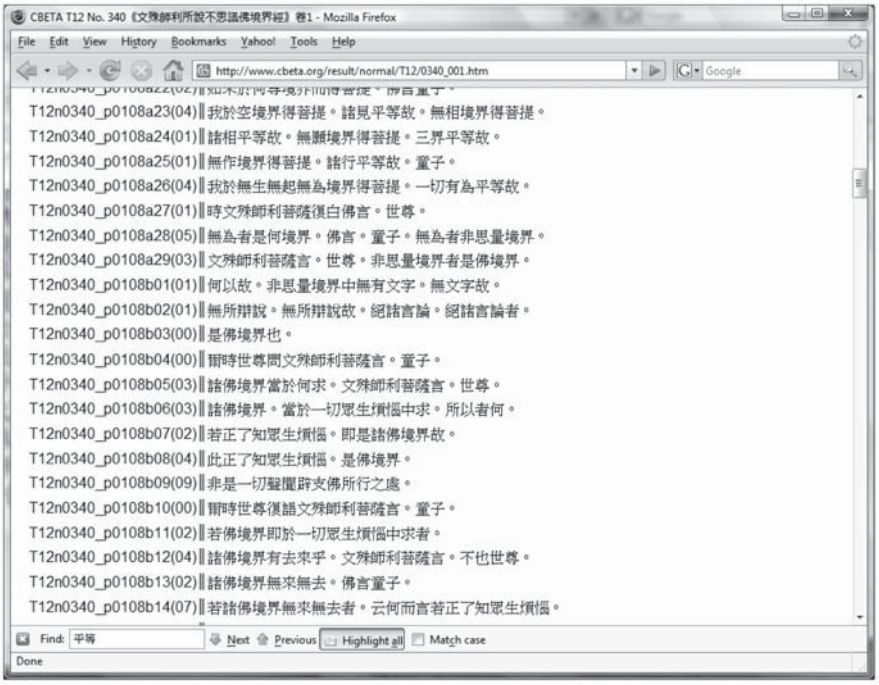
Another way of display is to use the Histogram procedure which shows the relative number of occurrences over time. Usually, this is applied to the appearance of words in sequence within a text. However, with our time tagged metadata, we can show the outline of occurrence over the centuries. Scholars using this window can quickly see the pattern of how often a term was used in any given time period.



While viewing the abstract "Blue Dot" display, the next window opens to show the scanned image of the rubbing made directly from the 13th century printing blocks of Hae-in Monastery. While it is possible to view the same text in the digital version, we have chosen to display the original which was copied to make a new version for the computer. With our interface, the characters on the scan can be highlighted since we have established a link between each position in the original and the "Blue Dots." The advantage for the user is that any questions about mistakes made in the input of the digital version can easily be checked against the original.

Here is the image of the scanned rubbing from the original.

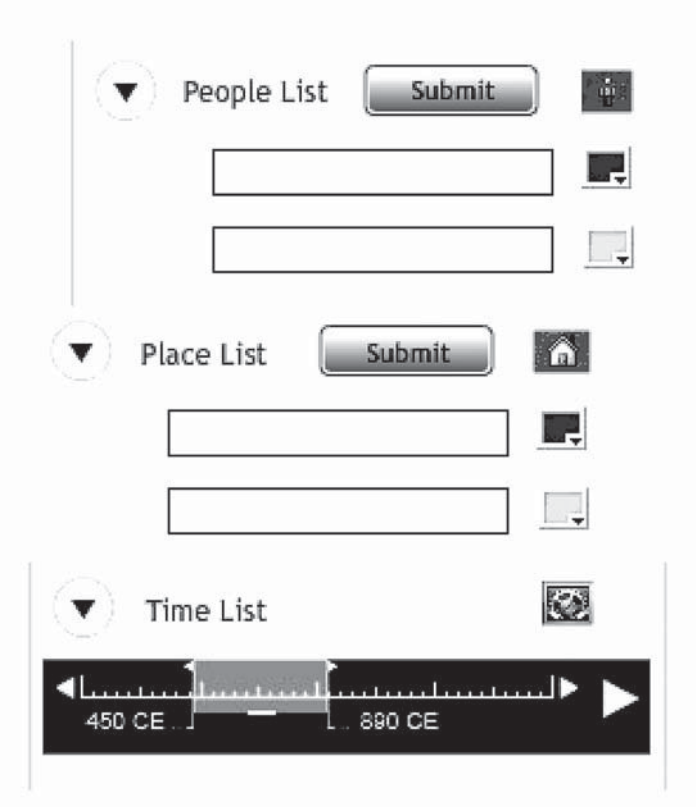
Here is the same page as it appears in the full text digital format. Users will be able to view this format as well as the scanned image shown above. The scanned image will be the default setting for display and the digital full text an on-demand element.



Our next widow is constructed to help give the user a context for the page being displayed in the image above. A Selection Window (shown in green) can be moved rapidly across multiple pages arranged as icons. The designated page appears above as a single page in the larger format of the main display box.

While we have shown all of the display panels opened on the draft interface, the user can control the number and type that are open at any time. The use for these buttons is enhanced by presenting an easily identified picture of the window..

In order to help the user define searches, we have the input box for the target word as described above. In addition, we have search fields for names of people associated with the text, place names, and the time of creation. In this way, the "Blue Dots" can be rearranged according to who translated the texts or the place where the translations took place or the time of the translation. With each new arrangement the "Red Dot" pattern changes to reflect the context of the search in terms of temporal and spatial elements.


People List
Submit
Place List
Submit
Time List
450 CE
890 CE