

Text Analysis and Pattern Detection: 3-D and Virtual Reality Environments

Introduction

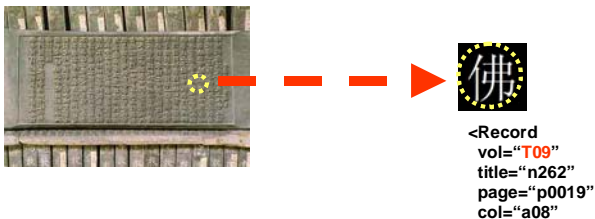
Our purpose is to explore the use of high dimensional visualization for analyzing text structure and patterns for scholars in the humanities. Digital Library programs are daily increasing the amount of material that is available but, humanities have yet to expand their research strategies to make full use of the potential of this technology. It is crucial that humanities scholars take their place among the technological strategists.

The Korean Buddhist Canon is the oldest complete set of the texts that make up the Buddhist canon for East Asia can be accessed on line through the Chinese Buddhist Electronic Text Association (CBETA) site in Taiwan. (<http://www.cbeta.org/index.htm>) With the help of colleagues in Japan, my catalogue of the canon was digitized 2001-04 and available for use. (http://www.hm.tyg.jp/~acmuller/descriptive_catalogue/) This proposal is the next major step in bringing innovative techniques to the canon.

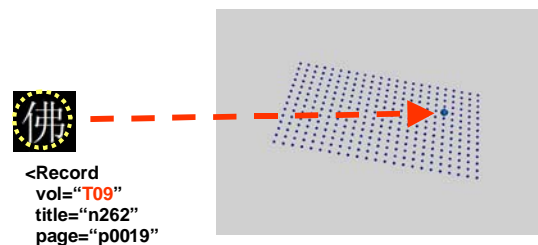
Project Description

The use of computers in text study has not yet embraced virtual reality and 3D visualization, an innovation that will require some positive examples of the value of the methodology before humanities scholars will accept it. A SGER can give proof of concept before proceeding with larger projects involving numerous researchers. We will:

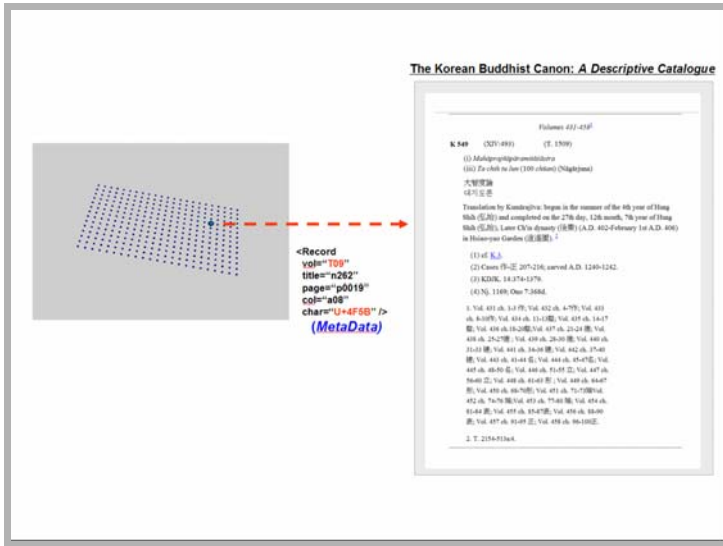
(1) Convert the glyphs that are displayed from the Chinese fonts into colored forms. Here we see a “page” of the material, a 13th century printing block, with 23 lines of 14 characters. All of the material appearing on the more than 83,000 blocks has been digitized in full text. Each of the characters was marked up with a Unicode designation , e.g.



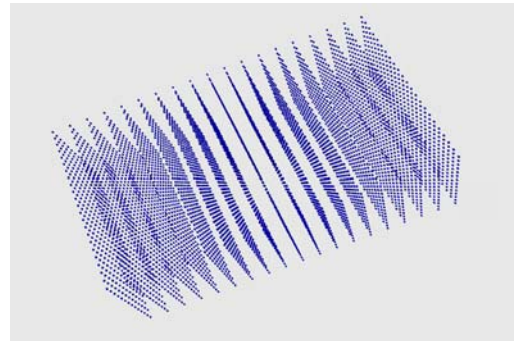
These millions of glyphs will be converted into an image (a blue dot in this case) that will have the same metadata as the glyph.



Because we have a digital catalogue of the set of texts, each blue dot can be linked to the appropriate description of the text in which it appears and so the Virtual Reality is closely tied to reference works and context.

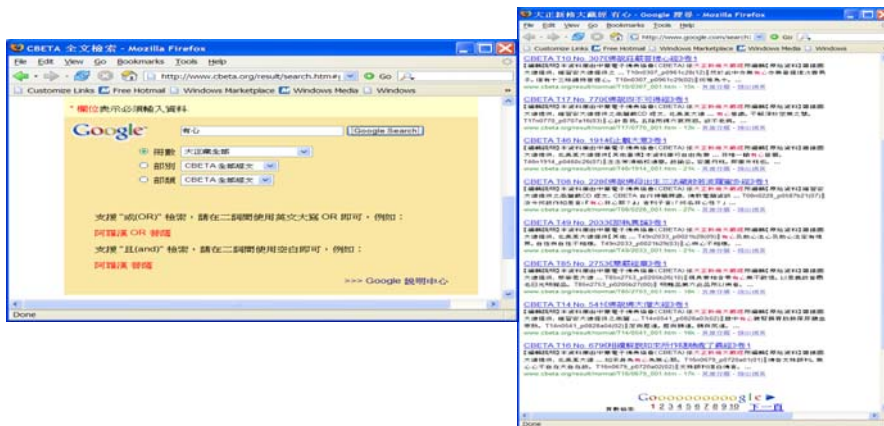


The blue dots representing the glyphs are then arranged in the order of appearance on “pages.” Each panel represents a page and each dot a glyph in its proper order.

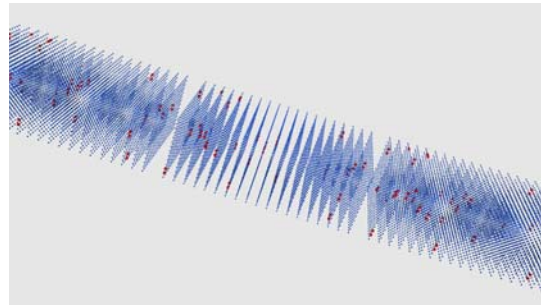


Searching for terms in the VR presentation

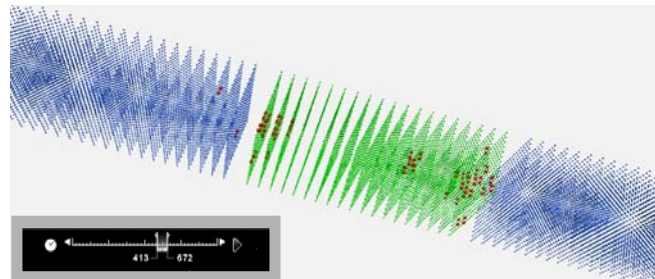
The search for patterns can start with a string search of the digital data. The result of such a search in VR will differ greatly from a conventional Google search display.



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.



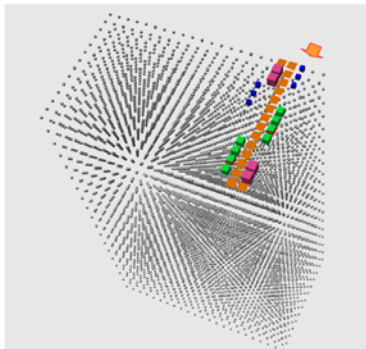
By using the context of the dots in the temporal sequence of translation dates, it may be seen that the word appeared regularly in the 3rd-5th century (Green section) and irregularly in the 9th-11th Blue section to the right.



Patterns of Structure in Content

In addition to being able to visualize word distribution and frequency, we want to explore more complex analyses, e.g. Chiastic structure. 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'. A “ring” construction 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 corresponds to the statement at the beginning. Anthropologist Mary Douglas, *Thinking in Circles: An Essay on Ring Composition* points out the need to make cross-cultural studies of this particular literary phenomenon.

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. 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. 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. We should pay close attention to the Orange dots (D) because they represent the theme or major concern.



Understanding this ring construction shows how we can use the abstracted dot structure to identify more complex patterns than a string search could.

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 (in brown) or as a thematic parallel e.g. blue, green or red.



There are many ways in which the VR display can exhibit our dots that represent the glyphs of the Buddhist canon. The approach can be used on a regular computer screen or it can be in a VR immersive theatre setting. Our hope will be to have such immersive settings in Australia, Korea, and perhaps Singapore, so that we experiment with teams of experts working together at the same time in the theatre, looking at the shifting imagery of searches, representations of structure, and distribution of occurrences. It is expected that this format will finally break and go beyond the bounds of our normal procedures.

Results Expected from the Project

The intellectual merit of the proposed activity is that using abstracted VR model in the search for word strings, clustering of terms, automatic analysis of Ring Construction, viewing results by time, creator, and place will provide textual scholarship with a new approach. The following are some of the ways that this strategy will permit novel results to permeate the search and retrieval process.

(1) Viewing the search results in the abstracted images allows the user to quickly spot visual patterns of occurrence of target words, rather than hundreds of lines of reports from the present Google search result, a needed innovation. Present techniques take the scholar many hours or even weeks to discern patterns of distribution when the search result contains hundreds of examples scattered across thousands of pages of text. A number of patterns of distribution are immediately available when the research result is an image.

Conclusion

The work proposed is a quite specific study using three components:

- (1) A large and important digital corpus;
- (2) A catalog containing descriptive metadata that are, in turn interoperable with other cyber infrastructure resources becoming available through our collaborative Religious Cultural Atlas of China and Himalaya (URL): place name gazetteers and map visualizations; biographical directories; lists of monasteries and monastic lineages etc.;
- (3) Virtual Reality techniques developed in other contexts for other purposes.

A specific project is necessary for proof-of-concept, but we see the broader impact resulting from the proposed activity as a strategic move with rich potential: the convergence of three development that have been heavily supported by Federal funders:

1. Large scale digitization in humanities, social sciences, sciences and engineering;
2. The development of metadata structures for description and as navigational infrastructure;
3. Virtual reality and high visualization data mining.

Our belief is that this small project could lead to new synergies – and not only in the humanities.