This is a report of the first phase of a project designed to create a new way of searching, retrieving, and analyzing digital texts. In the current funded project, we have made progress on data collection, data analysis, visualization and interface layout.

The figure below shows the 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 full digital 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 are 25 GBs in low resolution. Following the acquisition of the canonic material, we constructed our search engine. Based on the Suffix Array Technology, the search engine leverages our implementation from the collaborating group known as Chinese Buddhist Electronic Text Association (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. With the repository of data in place we developed the search engine, which is activated from the first window of the interface.

The first function of the search engine brings up every occurrence of the string 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 dynasty 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:

a. Multiple string search (allowing for cluster identification)
b. “String-Pair-in-Distance” (allowing for search where a second occurrence of the target may be distant by many lines in the text).
c. 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.
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. Additionally, 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.
We are still working to make this interface fully functional. Our greatest problem is that the current web technology to handle interactive 3-D images of millions of Blue Dots at once is still limiting. We are still trying to find a solution for this technical challenge.

**Report to scholarly community**

Power-point presentations of this information have been made to some of the potential audiences for use of the procedure in the following places:

1. Computing and Humanities Annual Conference  
   University of Illinois, Champaign-Urbana
2. Chicago Humanities Computing Seminar (Keynote)  
   Northwestern University, Evanston
3. Semantic Web Annual Conference  
   Busan, Korea (workshop)
4. University of Pittsburg Digital Library Forum (Keynote)  
5. Tripitaka Koreana Workshop  
   Seoul National University, Korea
6. ECAI Congress of Cultural Atlases  
   Academy of Sciences, Moscow, Russia
7. Fudan University Lecture  
   Shanghai, China
8. Vietnam Buddhist University Lecture  
   Ho Chi Minh, Vietnam
9. Ancient Buddhist Text Workshop  
   Hanazono University, Kyoto, Japan
10. Ecole francaise d’extreme orient Demonstration  
    Siem Reap, Cambodia
11. National Digital Archive Program Conference (Plenary)  
    Academia Sinica, Taiwan
12. Huafan University Lecture  
    Taiwan
13. Fo Guang University Lecture  
    Taiwan
14. United Nations Vesak Day Conference (Plenary)  
    Hanoi, Vietnam

The response from the presentations has indicated that scholars in a wide range of research see the potential of our approach in their own fields. The requests for beta testing include: Jane Austen project for review of patterns in her complete corpus, Supreme Court Case Law references, photo archive of 300,000 images at the Huntington Archive at Ohio State, Irish Journal Project of 500,000 pages scanned from 150 journals, and Vietnam Inscription Project of Ecole francaise d’extreme orient dealing with 50,000 images.

At the end of this first phase of research and development, we have a working model of the tool that can perform the tasks described in the funded proposal. During the work, our team has made major advances in terms of future plans for expanding the capabilities of the approach. We believe that it will be possible to incorporate many existing strategies for analysis and display as well as opening up innovative approaches for search and retrieval. While we have mainly worked with Chinese, early experiments
show that the imagery modeling can work with Roman script texts as well. The positive response from the many public demonstrations and lectures is an encouraging sign that the scholars in the Humanities are ready for this step beyond the current search and retrieval methodology. The progress made during the first phase is encouraging and we are more than ever aware of the need for expanding the development of this type of research tool.