Text-Image linking of Japanese historical documents: Sharing and exchanging data by using text-embedded image file

Okamoto, Takaaki
04c0004@sch.otani.ac.jp
Digital Humanities Center for Japanese Arts and Cultures, Ritsumeikan University

This paper will demonstrate the effectiveness of the linkage between textual data and image data in the field of medieval Japanese history where the research primarily uses text-based documents instead of photocopies and digital images.

1. Using Computers and Data for Historical Studies

The current state of the field of Japanese history requires that researchers (1) view more historical documents than in the past, (2) fully utilize the various types of information contained in these documents such as the form of a character, handwriting, type of paper etc., and (3) collaborate with others on inter-disciplinary projects instead of pursuing isolated projects in a single field. It is not possible to realize these goals through the individual efforts of researchers. A well-established digital environment becomes necessary. The most important and fundamental research tasks for a historian include: (1) looking up certain texts: what documents include these and where in the documents they can be found, and (2) examining the target documents. Given the nature of such research work, the utilization of computers is not as advanced as it should be. Without a fully computerized environment to enable easy access, the groundwork of identifying and reviewing documents would take considerable time and effort, and thus impede the advancement of research. If images and texts of documents are digitally connected, by searching keywords or sentences, the system will display the target texts directly as a part of the image, and users would also be able call up lists of entire images by inputting single characters

2. Semantic information and visual information of Historical Documents

In the field of history, most commonly circulated materials are published in traditional printed media such as reprints of texts. Historical documents that have not been reprinted are used only when they are significant for specific research projects. For historical studies, much of the emphasis is on the interpretation and analysis of primary texts, and reprints of primary material, in plain text form, are often used for such purposes.

Inevitably, certain information such as the original handwriting will be lost in the process of transforming the original into plain text. In addition, handwritten material published during the pre-modern period used numerous variants of Chinese characters (異体字) different from modern publishing standards. Many of the variations of old characters are not represented by computerized fonts and thus are often lost in the conversion into text files. Kunten (訓点), the punctuation marks to read Chinese text (Chinese classics, sutras translated to Chinese, etc.) in Japanese pronunciation and grammatical order, are represented by special symbols. For research using this type of document, simple text files are not enough. The primary material needs to be converted into image files.

In other words, historical documents contain semantic information that can be converted into text files as well as visual information that cannot be represented in text files.

3. Possibilities of Visual information

To use handwriting data as an example, historical documents often do not indicate the author's name. Even in cases when a signature is included, it can often be the person who authorized the document and not necessarily the actual penman. For this reason handwriting becomes an important piece of information to determine the composer of

the document. Graphology analysis can reveal answers to the following questions: (1) who manually composed this document, (2) why did this person write this document, and (3) what other documents have possibly been written by the same person? In conjunction with content analysis (interpretation), these additional aspects can further advance historical and diplomatic research. Introducing imagery analysis to research that has been centered on textual analysis will no doubt lead to new developments.

4. Computer Environment for Handling Historical Documents

Although researchers have long recognized the importance of visual information, limited research has been done in these directions. One main reason for this is the lack of a well-developed digital environment, which has made such work highly labor intensive.

In order to make searching digitized images just as convenient as searching digitized text, images need to be organized based on the questions of "what character is contained in what part of which document." It will also be necessary to provide easy to understand results that will highlight searching characters or text within the image. To enable such functions, text data needs to be correlated with image data through coordinates by each character. In other words, three types of input are needed to set up this coordinate system: (1) Which document text does this image correspond to? (2) What characters or words are included in the text? (3) Where in the image are the search words? Two of these inputs, the image data and the text data, are already in common use. If every individual character of the text can be linked to the image through a coordinate system, it will enable text searches within the image of a document.

5. Text-image Management Tool and File Sharing/Exchange via Portable Devices

Researchers need to build an easily accessible digital environment with the documents and images of the documents and would thus find it useful to have a tool that could indicate the position in the original documents for a given text of interest. We are currently developing a tool that processes image files of text documents in this way (Fig. 1). Using this tool, a researcher, by clicking on any character in the image, can create information about the character such as its position within the image. As shown in figure 2, the data created for each character including ID (automatically generated GUID), position in the text, coordinates, dimensions and so on are made into a simple text file that can be managed by external software applications like Excel.

Figure 1

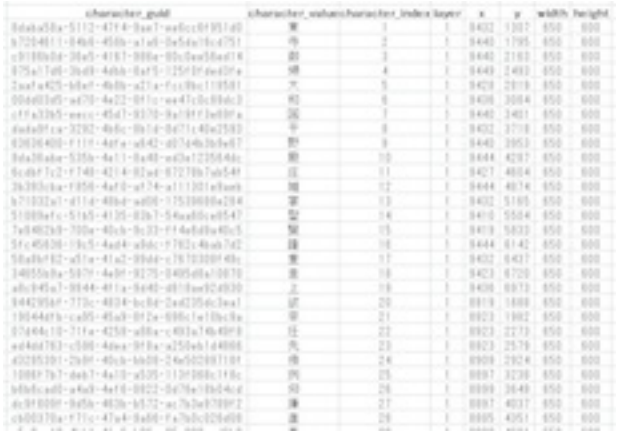

Figure 2

The program can generate reduced images for viewing in a browser and an HTML file containing information of the positions of the characters. When users search for a string of characters, a Javascript function will highlight the search results on the generated HTML file for display (Fig. 3).

Figure 3

The process of connecting these two sets of information cannot be automated and is dependent on manual input, which will inflict certain costs. Moreover, since the productivity of data processing by individuals is limited, group collaboration is essential. For the sake of simpler file sharing, this software thus creates one single file in TIFF format that combines the image data and the text data files. This enables the sharing of text data and image files through the exchange of just one file, an image file embedded with text data through a portable device. Moreover, on the receivers' end, it will be possible for all the image files of the individual characters and the HTML files needed for searching text strings to be generated from the one text-data embedded image file, eventually making all of these exchanges unnecessary.

6. Publishing on the Web

Although owners of the documents are often unwilling to publish the images of their documents on the Web, it is still necessary to develop a network friendly environment for those documents to be published. Such an environment consists of web servers, databases, and web applications that connect them. Since the system needs to be user-friendly for the researcher, there is a tool for importing image files with embedded textual data as mentioned previously by a simple drag-and-drop function. Once the file is dragged and dropped into the tool, it will generate the following: (1) an image for display on the Web (2) coordinate axes for this image (3) partitioning of the image for zooming and update temporary text/index data in the database server. Then the image file and textual data will be ready for publication on the web.

7. Conclusion

This method uses data collected and compiled by individual researchers on their personal computers. Data sharing is done through external portable devices. Data sharing via external portable devices may cause delays in distributing the data, prevent other collaborators from referring to updated information, and ultimately cause inconsistencies in data references. The inconvenience of external portable devices may lead to situations where the data are not updated at all on the computers of other collaborators. To avoid these situations it is important to take advantage of the online environment. However, sometimes owners do not want researchers to put images of their historical documents in the environment. This paper demonstrates how to deal with such situations.