A Study on the Digitalization of Xixia Historical Documents

Changqing LIU

Center for Lexicographical Studies, Guangdong University of Foreign Studies, Guangzhou, China

Keywords: Xixia historical documents, digitalization, Xixia characters, Xixia Database.

Abstract: As Xixia Studies make further progress in recent years, digitalization and textualization of Xixia characters and historical documents are of paramount importance. This project explores the specific approaches of digitalization and cyberalization of Xixia characters, aiming to create a comprehensive platform for Xixia studies at the international level, which enables webpage display in Xixia characters, fast search, updating and document display. Moreover, the project discusses the interdisciplinary approach involved by drawing on methods in computer science and history studies represented by studies on historical documents of ethnic minorities.

1. Introduction

The Western Xia (Xixia) Dynasty was a minority regime established by the Dangxiang minority in ancient China. It has gone through eight emperors during the past nearly two hundred years (1038-1227) of Xixia's existence. The Western Xia Dynasty left behind a large number of documents of the Xixia, which are of great value to the study of the Chinese history during the Western Xia Dynasty. The similarities between Xixia characters and Chinese characters are as follows: (a) they both belong to the ideographic writing system and Xixia characters are difficult to recognize, write and memorize; (b) they belong to square characters, and the strokes of Xixia characters also involve Chinese characters'points, horizontal, vertical, abbreviation, abbreviation, mention, etc. (c) they have similar word formation, 80 per cent of Xixia characters belong to synthesized ideographic characters; (d) there are Chinese calligraphies of the Kai; Xing; cursive scripts, and seal scripts both in Xixia character and Chinese. The differences between Xixia and Chinese characters are as follows: (a) The strokes of Xixia characters are complicated, and most of the characters are more than ten strokes; (b) Xixia characters have more oblique strokes and full corners in shape; (c) Xixia characters do not have the apparent radical system of Chinese characters; (d) Chinese characters remain many pictographs as it was developed from a pictograph in primitive society. However, Xixia characters were created only after the Dangxiang people entered feudal society, with very few pictographs left; (e) The grammatical structure and syntax of the Xixia character are different from those of the Chinese language, for example, the Chinese words "第一" (means "first" in English) was written as "一第〃", "开渠" (means "digging canal") was written as "渠开", "下雪" (means "snowing") was written as "雪下" and so on. Up to now, more than 6000 characters have been found and sorted out from Xixia historical documents.

In recent years, the study of Xixia has attracted significant attention from scholars at home and abroad, and great progress has been made in this field. A large number of documents and materials about Xixia have been photocopied and published [2]. At present, the study of Xixia characters and Xixia historical documents remain at the stage of traditional manual searching, which is time-consuming and highly laborious. The most critical literature works in Xixia academic circles are as follows: (a) The Documents of Heishui City Collected in Russia, published by Shanghai Ancient

Copyright © (2022) Francis Academic Press, UK DOI: 10.25236/edssr.2022.053

Books Publishing House, has released 11 volumes so far; (b) The Documents of Xixia Collected in China, published by Dunhuang Literature and Art Publishing House, has 20 volumes in total; (c) The Documents of Xixia Collected in Britain, published by Shanghai Ancient Books Publishing House, with five volumes in total. Among them, "The Documents of Heishui City Collected in Russia" boasts the most extensive collection of Xixia historical documents [3]. In addition, the French National Library also has some collections of Xixia historical documents, which Paul Pelliot discovered in Dunhuang Mogao Grottoes in March 1908. The number of documents in Xixia collected in France is less than those in the other three works mentioned above. Besides, there are also a tiny amount of Xixia historical documents in Japan.

The study of using the computer to process Xixia characters can be traced back to 1972. At that time, Grenstein of Denmark designed a computer coding scheme for Xixia characters, but it finally failed to manage to do it [4]. In the late 1990s, Professor Li Fanwen, a scholar focusing on the Xixia study, designed the Four-Corner number and Five-stroke typeface input method similar to Chinese features for the input of the Xixia character into the computer. Professor Li's Four-Corner input method is adopted in most of the Xixia character computer editing software in China. In the international field, scholars from Japan, Russia, the United States and the United Kingdom have done some research work in computer processing of Xixia characters. In 1996, the Japanese National Institute of Asian and African Languages and Culture developed the Xixia character library and typesetting system. The associate professor of this institute, Shintaro Arakawa, co-authored the *Xixia character Dictionary* with Russian expert Kochanov. The Institute of Linguistics and the Institute of Information Science of the Taiwan Academy of Central Studies of China began to develop the Xixia character Library in 1999 and successfully completed it in 2000. In 1997, Professor Li Fanwen and Japanese scholar Nakashima, using the typesetting system mentioned above, co-authored the book *"Research on Computer Processing of Miscellaneous Words in Xixia character"* using the typesetting system. In November 1999, Ma Xirong and Liu Changqing of Ningxia University designed the software of Xia-Chinese Character Processing and Electronic Dictionary. This software is a single-machine version of Windows application software that arranges, annotates and defines the Xixia characters according to the Four-Corner number and the sequential number checking method. In 2005, Jing Yongshi and Jia Changye developed a Xixia character input system based on the square code system. This system established the Xixia character library through split joint and modification of Chinese characters. Scholars of the Institute of Ancient Documents in Western Regions of Tongji University, including Ye Jianxiong and Shandi have designed an optimized system structure for the Xixia phonological thematic database by means of computational linguistics and established a document database structure oriented to the Xixia phonology. In 2011, the Xixia Research Institute of Ningxia University developed the Xixia intelligent input method and the network platform software for the textualization of Xixia historical documents [8]. In recent years, the research achievements on Xixia character digitization mainly focus on the computer extraction and recognition of Xixia characters [9-10].

2. Xixia Historical Document Database

The overall structure of the Xixia historical document database covers architecture, art, literature, academic research and other aspects of Xixia. Specific research to construct this database can be carried out from phonology, linguistics, the corpus of Xixia character, and the photo gallery database of Xixia character cultural relics. Before constructing the Xixia historical document database, the structure of the Xixia literature database should be established according to the contents of the documents. Through the establishment of a database structure, we can further sort out the specific

content of the Xixia historical documents database.The Xixia historical document database can be divided into five levels. The first level is the general classification of whole documents, and the second level is the subclassification under each classification. For example, Xixia architecture can be divided into five main categories, of which "mausoleum", situated in the secondary classification, can be further divided into "emperor mausoleum" and "family mausoleum". The third level classification is a further refinement of the second level classification. According to the written language, the Xixia literature can be divided into "Xixia character", "Chinese language", and minority language. Among them, "Xixia character" can be further divided into "secular literature" and "religious literature" in the third level. The "secular literature" can then be further subdivided into the fourth level. The final structure of the Xixia historical document database can be divided into five levels.

2.1. Retrieval System of Xixia Historical Document Database

The retrieval system of the Xixia historical document database is a software developed for the convenience of scholars on Xixia study. The system enables the full-text fuzzy retrieval of Xixia language and Chinese keywords. The retrieval data includes seven attributes: the number of Xixia historical documents, the name of Xixia historical documents, the source of Xixia historical documents, the content of Xixia historical documents, the Chinese translation (the contrasting Chinese content), the notes and the original images of Xixia historical documents. Compared with the traditional manual document search tool, it can find the literature content related to keywords quickly and accurately. The contents containing keywords are highlighted and hyperlinked with the original document images so that users can make a comparison of the recorded text with the original document. Figure 1 is the interface diagram of retrieval tools for the Xixia historical document database.

In Figure 1, the list on the left column shows the catalogues of all Xixia historical documents in the database. When a user chooses one of them, he can see the detailed information related to the document in the right edit box. As shown in Figure 1 as an example, when selecting the document as*佛说佛母出生三法藏般若波罗蜜多经*, the upper right edit box shows the specific origin of the document as "Britain collected Or. 12380-3392 Chinese translation of*佛说佛母出生三法藏般若波罗蜜多经*, Volume 25. The middle editing box shows part of the original Xixia text of this document, and the lower editing box shows its Chinese translation. At present, this retrieval system has entered 12,500 items in Xixia characters, mainly including Xixia characters, dictionaries, some Tangut Sutras and Chinese translation versions of Xia documents, etc.

The retrieval function of the system enables users to retrieve the keywords in Xixia characters or Chinese. Firstly, users need to select the keyword type, that is, the drop-down button marked "Xixia character" in the upper part of Figure 1. If we need to find the relevant content of the Chinese translation, they can choose the "Chinese language" in the drop-down list. If they need to find Xixia historical documents in the Xixia character, they can choose the "Xixia character". What the users need to do is to enter the content into the search keyword input box, and then a retrieving process can be completed by clicking the "search" button. For example, if a user wants to inquire information about "父子"("father and son" in English) in Xia's translation of *Mencius*, he or she needs to enter the keyword "噬蚀" in Xixia character, and the retrieval result can be shown as in Figure 2.

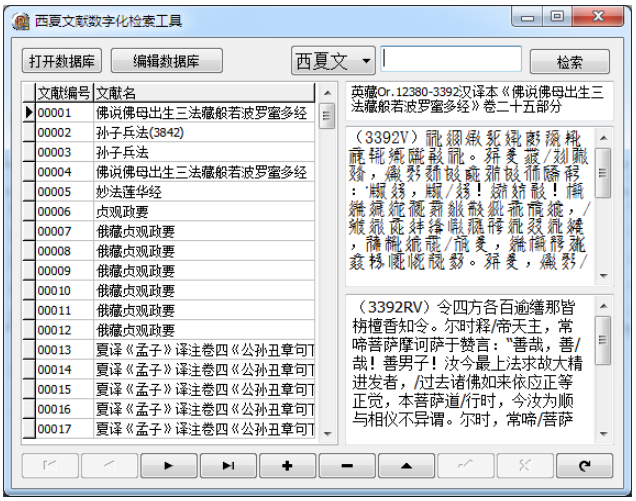

Figure 1 Search Tool for Xixia Historical Documents Database.

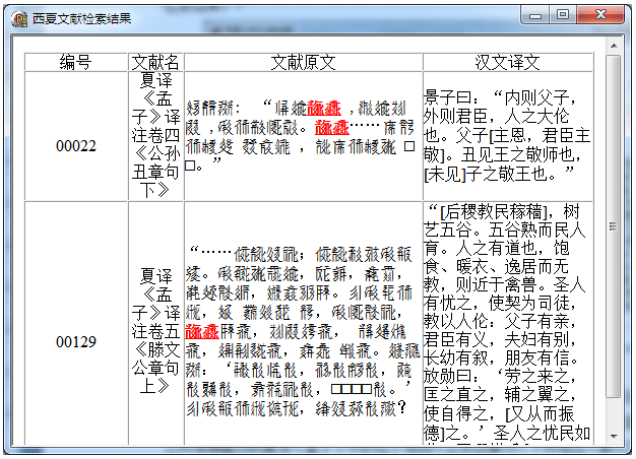

Figure 2 Search Results of "噬蚀".

"噬蚀" in the picture literally means "son-father" in Xixia character, which is translated into "father-son" in Chinese in a reversed order. Translators in Xixia character often invert the word order of a kind of nouns consisting of juxtapositions in Chinese, sometimes even the order of some coordinate-compound sentence. The retrieval system can accurately find two original Xixia texts in Xia's translation of *Mencius*, which are located in Volume 4, *The latter Chapter of Kung Sunchou*, and Volume 5, *The Former Chapter of Tengwen Official*. The Chinese translation is also shown in the right box, which is convenient for researchers to interpret. The retrieval method can significantly improve the speed of Xixia historical document searching as well as the retrieval accuracy. Similar methods can be used to find the Xixia historical document containing the Chinese words "曹操". Fig. 3 is a diagram for the search result of the corresponding words "蕫繵" in Xixia character.

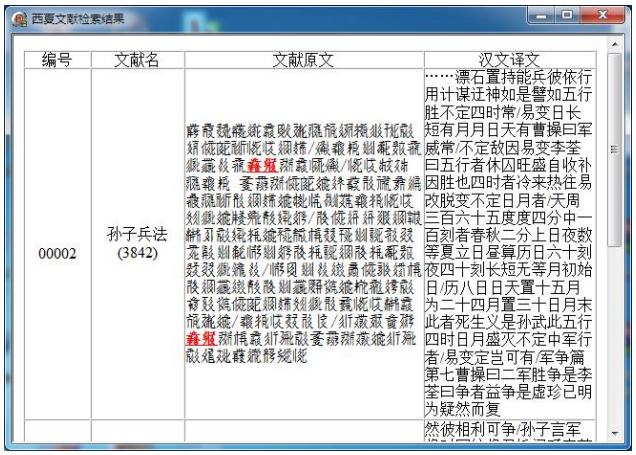

Figure 3 Search Results of "**蕤繾**".

Clicking on the red link in Figure 3, users can get the original document picture in Xixia character, as shown in Figure 4, which is a quick way to obtain the original document picture. Compared with the original manual method of consulting cards and documents, the retrieval system can save much time, providing accurate electronic materials for Xixia studies in a fast way. The original documents and images shown above are stored in the network server. What users need to do is just to connect to the Internet, and then they can use the electronic version of the original documents and materials of Xixia historical documents.

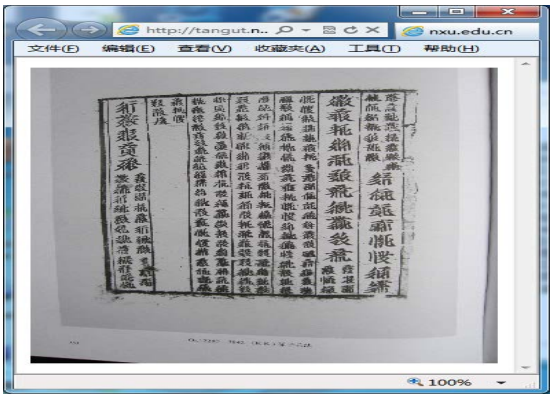

Figure 4 Original Image of "**蕤繾**".

2.2. An Example of Searching Xixia Character Dictionary Database

In order to illustrate the validity of this method, nine kinds of dictionaries and related secular documents in Xixia character are selected as the retrieval collection, and five commonly used concepts listed in the following table are searched in this full retrieval collection text using this method. The search results are shown in Table 1.

Table 1 Search Results of Xixia Historical Documents.

Keywords	Counts	Results（top 5）				
姓氏	288	**摆**	**疥榜**	**眛**	**睋**	**睗**
草	85	**步**	**部**	**砑**	**矶**	**矽**
菜	24	**粿布**	**裬**	**福**	**袱**	**跟**
白	13	**褖**	**碉**	**端妒**	**篁**	**溉**

病	24	癵	传	耔	耘	稠

Table 1 shows five common words such as "姓氏(surname)", "草(grass)", "菜(vegetable)", "白(white)" and "病(disease)". The retrieval results all include the content containing these five common conceptual words mentioned above in the Xixia character. As long as one of the five commonly used words is included in the search results, it can be considered as a valid tuple. By means of fuzzy matching retrieval, tuples containing keywords can be retrieved to the maximum extent in the database. In this case, 288 items of "姓氏" contain Xixia character entries of "白" as a surname. However, there are also Xixia character entries of "白" surname when searching for "白" as colour. Therefore, in this case, the 15 records of "白" should be filtered manually, and 13 valid records remain after the filtering process. The main reason for this kind of search results intersection is that there lacks "semantic understanding" in this retrieval process. Consequently, the full-text retrieval with "semantic understanding" needs to make further study. Then retrieval results shown above are compared with those gained through the manual retrieval commonly used by scholars on Xixia study. Take the most commonly used Xia-Chinese Character Dictionary (2008 new edition) as an example, in this dictionary, there is only one index for the index entries of "姓(also means surname)" or "姓氏", which is far less than the retrieval results in this database. For the index entries of "草", "草", "草名" and "草木", there are 64 indexes which are less than retrieval results in this database of 85 entries. As for the index entries related to "白", there are six indexes that are less than retrieval results in this database of 13. Finally, there are nine index entries related to"病", which are less than the retrieval results of 24 entries obtained from this database. Compared with the traditional manual retrieval, Using Xixia historical document database to make retrieval can obtain more comprehensive data. The tests above are only conducted in part of the Xixia historical documents. If a full-text retrieval is conducted in the complete Xixia historical documents database, the retrieval results will be more abundant.

3. Conclusion

The study on digitalization of Xixia historical documents is of great significance to promote the study of ancient characters and documents of other ethnic minorities in China, such as Qidan and Nvzhen. Combining with the network platform, the processing of the Xixia character and its historical document can facilitate academic exchanges and study activities between Chinese and foreign scholars. Through the integration with the digital resources of the Xixia character, the Xixia research platform under the network can be built to realize the digital preservation and resource sharing and utilization of the Xixia character and Xixia historical documents. As an essential part of Chinese history and culture, the study on the digitalization of Xixia historical documents will have a profound impact on Xixia studies. With the characteristic of fast information updating, this platform will be a powerful tool for people to study and understand Xixia culture. The study on the digitalization of the Xixia character and its historical documents under the network will blaze a trail for the research and popularization of Xixia culture.

Acknowledgement

Supported by Key Research Projects Funding of Colleges and Universities in Guangdong province in 2018 (Project No.2018WZDXM011).

References

[1] Li, F. Comparative Study on tongyin & tongyi in Tangut. Collection of Tangut Studies. Shanghai: Shanghai guji Press, 2012: 161-167.

[2] Liu, C., Du, J. On processing Xixia characters and Xixia historical documents on Internet. Social Sciences in Ningxia, 2008, 5:113-115.

[3] Du, J. An introduction for Xixia literature Collected in China. Xixia Studies. Ningxia: Ningxia People's Press, 2006.

[4] Nie, H. A study of Tangut characters in the 20th century. Xixia Studies in the 20th Century. Ningxia: Ningxia People's Press, 2004: 112-124.

[5] Li, F. Research on Xixia dictionary on edit, sijiaohaoma and input method. Social Sciences in Ningxia, 4, 1997.

[6] Ma, X. Xixia Word prcessing & electronic dictionary.Beijing:Tsinghua Universtiy Press, 1999.

[7] Liu, C. Present situation and prospect of xixia computer digitization—Xixia Studies. Shanghai: Shanghai Guji Press, 2011.

[8] Liu, C. Design and Implementation of Online Xixia-Chinese Electronic Dictionary. Journal of Ningxia University (Natural Science Edition), 2011, 32:349-352.

[9] Meng, Y., Zhang, X., Yang, X. Application of feature extraction and matching based on ASM in OCR. Journal of Guangxi University (Nat Sci Ed), 2017, 42: 2183-2190.

[10] Lin, Y., Bi, B., Sun, F. etc. Identificaiton method of movable type printing of Tangut characters based on image registration. Journal of Lanzhou University of Technology, 2016, 42:97-101.