

Penn Libraries

Scholarship at Penn Libraries

University of Pennsylvania

Year 2007

Approaching pre-modern China through
the computer: the benefits and risks of
using electronic resources in sinological
research

Jidong Yang

University of Pennsylvania, jdyang@pobox.upenn.edu

Postprint version. The author's paper was presented at the *Annual Meeting of the Association for Asian Studies*, March 2007.

This paper is posted at ScholarlyCommons.
http://repository.upenn.edu/library_papers/29

Approaching Pre-Modern China through the Computer: The Benefits and Risks of Using Electronic Resources in Sinological Research¹

Jidong Yang

University of Pennsylvania

Introduction

Around the middle of the 1990s, Chinese Studies both in China and around the world irreversibly entered the digital era. Only a decade later, the means and resources for studying pre-modern Chinese civilization have dramatically changed. Online library catalogs, Internet-based bibliographies and indexes, electronic journals and books, and full-text databases, many of which were beyond the wildest imagination of scholars of the previous generation, have now become indispensable tools for daily research. Especially noticeable is the large-scale digitization of pre-modern Chinese texts which has, literally, revolutionized the information-seeking behavior and research process of today's sinologists. The improvement of Chinese OCR (optical character recognition) technology has made it possible to scan and index traditional Chinese books at a manufacturing speed. By the time we gather for this conference, the vast majority of pre-Song written works, except perhaps Daoist scriptures and archaeologically excavated manuscripts, have been digitized at least once. And it is just a matter of time until most of the post-Tang literature will be available in an electronic format.

These new types of resources for Chinese Studies have been the subject of scholarly attention for quite a few years. But so far most published writings on the topic concentrate on technological issues related to hardware, software, character encoding, access, and so forth. Discussions of the new resources' impact on sinological research can be heard from time to time, but most commentaries have been very general and

¹ This is a draft of my presentation to the Annual Meeting of the Association for Asian Studies in Boston, March 22-25, 2007.

excessively positive. In fact, like all other revolutionary media of information that have appeared in history, such as paper and printing in early and medieval China, the impact of digitized texts on contemporary scholarship and academic culture is far more complex and needs to be examined more carefully. In my speech today, I would like to focus on full-text databases of pre-modern Chinese written works and discuss their impact, both positive and negative, from a user's perspective.

Resources examined

My discussion is based on the examination of a number of full-text databases, especially the following three:

1) *Scripta Sinica* (*Zhongyang yanjiu yuan Hanji dianzi wenxian* 中央研究院漢籍電子文獻). A pioneer in the digitization of the pre-modern Chinese literature, this database was initiated as early as the 1980s and remained the only resource of this kind available to most scholars until the late 1990s. Full access to the database is limited to institutions and individuals in Taiwan as well as a few selected overseas institutions. A small part, which contains some of the most important sources such as the Confucian classics and the twenty five dynastic histories, is open to the general public through the Web. All texts in the database have been input manually and undergone proofreading multiple times.

2) Electronic *Siku quanshu* 四庫全書. This database was produced by a Hong Kong company and is the digitized version of the Wenyuan Ge 文淵閣 copy of the *Siku quanshu*, which is now located in Taiwan and has been photo-reprinted a few times since the early 1980s. Unlike the *Scripta Sinica*, the electronic *Siku quanshu* was made by scanning and indexing the original hand-written copy.

3) *Guoxue baodian* 國學寶典 database. Still unknown to most North American institutions and researchers, this database was first produced by a company in Beijing on CD-ROMs during the late 1990s and soon became the most popular electronic resource for mainland Chinese scholars. After a number of upgrades, the database went online about a year ago and is now accessible through either institutional or individual subscriptions. Its contents are apparently from a variety of sources: some are digitized from modern punctuated editions through manual input, whereas others are the results of scanning pre-modern editions without punctuation.

The benefits of digital resources

First of all, I would like to discuss the benefits of these full-text databases for studies of pre-modern Chinese civilization from a historical perspective.

From the point of view of information science, I think the history of Chinese Studies can be roughly divided into three major ages. The first or pre-modern age lasted for thousands of years until the 1920s. During this long period, information retrieval was largely based on personal memory and reading, although various reference and citation books did exist. The importance of a good memory is evident throughout a typical Chinese scholar's lifetime. At the first stage of primary education, he had to memorize character books such as the *Qianzi wen* 千字文. Later he was trained to recite major Confucian classics and well-known literary writings. When he grew up and began to pursue social success, he would find it impossible to pass the civil service examination without a good memory of the classical literature. In order to make himself better known to the elite class, he had to attend various parties and improvise verses by making prompt allusions (*diangu* 典故) to traditional works. The quality of his personal and official writings, such as political essays, memorials to the emperor, monographs on philosophy and history, and funerary texts for family members, was thought to be closely related to his ability to make appropriate references to the literature of previous ages. Even in later and modern times, those leading scholars in the studies of traditional China are all characterized by their familiarity with the primary sources in the field, and by their ability to cite passages from pre-modern works without even opening the book. A good example in this respect is the legend of the late historian Chen Yinke 陳寅恪. It is well-known among Chinese scholars that he was able to teach and research even after he became blind.

I would call the second age in the history of Chinese studies the age of print indexes. During the 1920-30s and under the influences of Western social sciences, the so-called "indexing movement" (*suoyin yundong* 索引運動) took place in China's academic world. The Harvard-Yenching Institute in Beijing (Beijing) and later the Centre Franco-Chinois d'Études Sinologiques published indexes to a number of pre-modern Chinese works, most of which dated from the pre-Song period. From the 1950s-90s, scholars in mainland

China, Taiwan, Hong Kong, Japan, France, and other places around the world compiled more indexes to traditional Chinese books. The most famous ones included the personal name and place name indexes to the 25 dynastic histories. Unlike scholars in traditional China, most sinologists born in the 20th century were never trained to memorize Confucian classics and other pre-modern texts except for a small number of short literary writings. Sinological research, therefore, became more and more dependent on the indexes, which to a certain extent determined both the achievements and limits of the discipline. While bringing unprecedented convenience to scholars growing up under the modern educational system, print indexes fell far short of providing an ideal method of retrieving information from pre-modern texts for at least two reasons. First, even after 60 years of painstaking compilation, published indexes had covered only a small portion of the entire literature of pre-modern China. And second, most of the print indexes are only for keywords or proper names, rather than the full text.

We are now ten years into the third era or digital age in Chinese studies. For only a decade, the development and improvement of full-text databases have allowed the younger generation of sinologists to compete with and even surpass their knowledgeable predecessors in both the speed and effectiveness of information retrieval. With the databases, exhaustive searches in a single or across multiple primary sources have become significantly easier and can be done in a few seconds, and the results are often more reliable than using manually compiled print indexes, not to mention personal memory. Especially useful are those databases that support advanced search by applying Boolean operators such as “and,” “or,” and “not.” Taking my own experience as an example, several years ago, I read in a newspaper that most ancient Chinese tombs were robbed at least once in history and thus left nothing valuable to modern archaeologists. I was interested in this phenomenon and wanted to get an idea how popular tomb robbing was in medieval China. Since I did not have enough time to go through all relevant primary sources for the topic, I decided to do a search in the *Scripta Sinica* by inputting “發&墓” (“&” stands for the operator “and”) as the search term and limiting the search to the *Jiu Tangshu* 舊唐書 (*Old History of the Tang*). The database quickly brought up all passages in the *Jiu Tangshu* that contained both characters. After browsing the 33 results, I found some valuable and interesting passages such as this one:

新校本舊唐書/列傳/卷七十二列傳第二十二/虞世南

自古及今，未有不亡之國，無有不發之墓。

It is evident from this memorial presented by Yu Shinan to the early Tang court that tomb robbing was already a widespread crime no later than Yu's time. More telling is this account from the biography of Guo Ziyi 郭子儀, a famous Tang general credited with saving the dynasty from the An Lushan rebellion:

新校本舊唐書/列傳/卷一百二十列傳第七十/郭子儀

大曆.....二年.....十二月，盜發子儀父墓，捕盜未獲。

Thus even the grave of Guo's father was robbed during Guo's lifetime. We can also find:

新校本舊唐書/本紀/卷十五本紀第十五/憲宗下/元和十四年

二月.....乙卯，敕淄青行營諸軍，所至收下城邑，不得妄行傷殺，及焚燒廬舍，掠奪民財，開發墳墓。

This is an imperial edict issued in the 14th Yuanhe Year (819) when the Tang army was fighting a war against some rebellious warlords in the east. Apparently, even the empire's troops were often involved in this highly profitable crime.

If did not have access to a full-text database, my attempt to solve this kind of research question would seem hopeless despite my knowledge of Tang history. However, with the help of electronic resources, scholars can now perform research tasks that were impossible or, at the very least, would have taken much longer in the pre-digital age.

Risks of using electronic resources

Nevertheless, our discussion of digital resources should not be limited to their benefits. In an age when, whether we like or not, all pre-modern Chinese texts are either being digitized or are on the waiting list, special attention to the shortcomings and limitations of the new resources is perhaps even more important to the health of the discipline. In my view, today's sinologists should be cognizant of at least two types of disadvantages of full-text databases, one is the short-term or immediate risks that occur in the daily use of such resources, and the other is the possible negative impact of digitization on Chinese Studies that may only become apparent in the long term.

The short-term problems include that of edition. As we know, a great many pre-modern Chinese works have more than one edition. Differences between various editions

of one source are sometimes significant enough to change the interpretation of the text, and in other cases are even associated with major intellectual and cultural transitions. For instance, the massive-scale replacement of the character *hu* 胡 (barbarian) in the Buddhist literature during the Sui-Tang period reflected a changed understanding of the nature of the religion. When compiling the *Siku quanshu*, the Qing government purposely altered many texts thought to be harmful to dynastic rule. By closely comparing and examining different editions of the same text, therefore, one may sometimes discover important clues to China's past. That is why certain knowledge of pre-modern Chinese bibliography has long been considered a required quality for a sinologist. However, when digitizing a primary source available in various editions, so far most database companies will choose only one edition and disregard all others in order to reap quick profits. As sinological research becomes more and more dependent on electronic resources, the lack of different editions will likely cause a simplified understanding of important aspects of traditional Chinese civilization. Here I also have to point out the fact that the *Siku quanshu*, the electronic version of which is perhaps the most popular research tool for today's students of pre-modern China, is well known as a low-quality edition for many primary sources. Besides, there are huge differences even among the seven copies of the *Siku quanshu*. A recent study shows that as many as 60% of the texts in the Wenlan Ge 文瀾閣 copy, which has not been digitized yet, are somewhat different from their counterparts in the Wenyuan Ge copy.

Related to the issue of edition is the undesired impact of full-text databases on textual studies. Although pre-modern Chinese works were written by authors who passed away a long time ago, many of them—those dating from the early periods in particular—are still in an “unstable condition.” That is to say, newly discovered materials, such as archaeologically excavated manuscripts and epitaphs, often make revision of old print editions necessary. Also, for those punctuated and collated editions (*jiaozhu ben* 校注本) published in modern times, even though they are generally considered the best editions of pre-modern texts, we often hear criticism about the accuracy of the punctuation. This tradition of textual studies has long been a driving force for sinology, but its chance of survival in the digital age may be questionable, because the convenience brought by full-text databases will likely decrease scholarly interest in print editions and manuscripts,

namely the foundation of textual studies.

We should also pay attention to the limitations of current digitizing technology. Due to the massive amount of literature of pre-modern China, it is financially unsustainable to make full-text databases solely by means of manual input and human proofreading. Although the manually made *Scripta Sinica* is widely praised for its higher quality, databases manufactured with OCR technology such as the electronic *Siku quanshu* and *Zhongguo jiben guji ku* 中國基本古籍庫 (www.cn-classics.com/chaoshi/index.html) are much larger in size and are apparently the mainstream. However, OCR technology has a known issue of accuracy. Even in the digitization of Western language materials in good condition, none of the OCR software providers can claim a 100% accuracy rate (a 99.9% accuracy rate still means an average of one error per page). So far I have not seen any published non-commercial evaluation of the accuracy rate of Chinese OCR software, but we have legitimate reasons to suspect that it is more difficult to achieve a high level of accuracy in indexing Chinese characters than Latin scripts, and in digitizing ancient materials than modern texts.

The quality of Chinese full-text databases is negatively influenced not only by the lower rate of accuracy, but also by the lack of enough characters in all three major coding systems: GB, Big5, and Unicode. Even for characters that are already included in these systems, they are not fully covered by any input software. Some databases made in mainland China, such as the *Guoxue baodian*, are only available in the simplified script which is, of course, not the script used by original pre-modern Chinese editions. All these technical problems, needless to say, reduce the value of full-text databases to various degrees. A simple search in the electronic *Siku quanshu* and the *Guoxue baodian* database by the character 曩 (Zhao, Tang Empress Wu Zetian's official name), for example, can reveal many problems.

But perhaps the biggest risk in machine-based sinological research is that the seemingly powerful search function provided by full-text databases may sometimes mislead the user into a premature or even false satisfaction over the search results. I can exemplify this risk with a real research case that I know. A graduate student writing a term paper about Kumārajīva, the famous Central Asian Buddhist monk and translator who traveled to China during the Sixteen States period, once did a search in the *Scripta*

Sinica version of the *Gaoseng zhuan* 高僧傳 (*Biographies of Eminent Monks*; also known as *Liang Gaoseng zhuan* 梁高僧傳) by the standard Chinese transliteration of the monk's name, Jiumoluoshi 鳩摩羅什, and found a total of six paragraphs. The student was satisfied by the results, but in fact what he found was a far cry from the full retrieval of relevant information. Kumārajīva is mentioned in the *Gaoseng zhuan* by a variety of appellations, such as Jiumoluoqipo 鳩摩羅耆婆, Shi 什, Shigong 什公, Shishi 什師, Tongshou 童壽, and Luoshi 羅什, all of which can be found in the best print index to this primary source, i.e. *Ryō kōsō den sakuin* 梁高僧傳索引, compiled by the famous Japanese scholar of Buddhist history Makita Tairyō 牧田諦亮. The student's search by 鳩摩羅什 alone, therefore, yielded only a small portion of the information concerning the monk. This example also shows that, in certain circumstances, full-text digital databases are actually less helpful than old-fashioned print indexes, because the latter were mostly compiled by known scholars who were very familiar with the primary sources in their fields and were able to establish cross-references between interrelated names and terms. In the future, even when every sinologist can enjoy easy access to all digital databases, these print indexes should still remain on reference shelves.

In addition to the short-term risks discussed above, I would like to briefly raise the issue of the long-term negative impact of digital resources on sinology. As we know, one can search for information, but must read to acquire knowledge. The convenience in searching brought by full-text databases, however, can decrease one's interest in reading books. I have to admit that I myself now suffer from the loss of interest in reading original primary sources. When I was majoring in ancient Chinese history at Beijing University in the 1980s, I was advised, or more precisely, commanded by my professor to read the *Zizhi tongjian* 資治通鑿 from the first to the last page. This reading experience, although quite painstaking and time-consuming, has proved important for my research even in the digital age. For instance, through the reading, I came to know that the geographic name Dunhuang was often written as 燉煌 in Han-Tang sources. So when searching in full-text databases for information about medieval Dunhuang, I know to input both 燉煌 and 敦煌, and will also try Shazhou 沙州, the more popular name for Dunhuang in the Tang-Song period. Unfortunately, my habit of reading primary sources did not last after the arrival of the digital age. Since then my knowledge of medieval

Chinese history and culture has been frozen at the pre-digital level. I believe that this lesson of mine is quite universal among Chinese Studies students both in China and around the world. As more and more pre-modern Chinese texts are available electronically and can be searched without a thorough reading, how to maintain self-discipline in careful reading has become a challenge to all of us. Failure to meet this challenge may cause a number of serious problems, such as a declining ability to understand pre-modern Chinese language and terminology.

Conclusion and suggestions

Having listed so many risks in using electronic resources for Chinese Studies, I hope I have not conveyed the wrong message that full-text databases are monsters. Overall, the new resources surpass old research tools by far in both data coverage and search function. That is why they are welcomed by most researchers. There is no way to turn back from the transition to digital sinology. My discussion of electronic resources' negative attributes are aimed at improving them, not accusing them of destroying sinology. To enhance the capability and reduce the risks of full-text databases, I think both users and database companies can make contributions.

As users, we must be aware of both the benefits and limitations of digital resources. Scholars should be more actively involved in exploring and developing effective search strategies for full-text databases. Graduate-level courses in Chinese studies methodology and Chinese bibliography should be frequently updated according to the latest trends in the digital world. To offset the potential long-term damage of digitization on future scholarship, stronger requirements for reading primary sources and relevant training programs may have to be established for coming generations of sinologists.

I am confident that the digitization of important pre-modern Chinese texts will be extended to different editions as the demand from researchers increases. The computer's language processing capability will continue to improve. As competition grows, the manufacturers and vendors of Chinese databases will realize that it is actually more commercially profitable to make their products more scholar-friendly and research-oriented. Finally, I would like to make an appeal to digitizing companies to work closely with scholars and design "smart databases" by creating cross references among

interrelated names and terms, so that when the user searches by a proper name, he/she will be able to retrieve all variant forms of the name.