10-1-1991

Twenty-Five Dynastic Histories Full Text Retrieval Database at the University of Washington

Yeen-mei Wu

Follow this and additional works at: https://scholarsarchive.byu.edu/jeal

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/jeal/vol1991/iss94/5

This Article is brought to you for free and open access by the All Journals at BYU ScholarsArchive. It has been accepted for inclusion in Journal of East Asian Libraries by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
The Twenty-Five Dynastic Histories Full Text Retrieval Data Base was initiated and actualized through the collaboration of the Institute of History and Philology, Academia Sinica (IHPAS) and the Computing Center of Academia Sinica (CCAS) for the purpose of creating a data base that includes documents from the twenty-five Chinese dynastic histories. IHPAS prepared the text and CCAS was responsible for input, quality control, etc.

The texts for this data base were taken from the Ting-wen shu chü's *Hsin chiao pen erh shih wu shih* (Taipei: 1986-87) which is identical to the edition published by Chung-hua shu chü (Peking: 1975-77). IHPAS has carefully reviewed this edition and made 135 textual revisions based on information from other authoritative editions. The total number of characters entered into the data base is 56,500,000, making it the largest Chinese data base in existence. It was developed from July 1984 to August 1990. One of the first Chinese data bases to be developed by Academia Sinica, the project was under the direction of Dr. C. C. Hsieh, then director of CCAS. The functions of the retrieval system are a description of the document structure, the ability to freely retrieve by any text term ("free text term"), a system-created table of contents, the ability to retrieve any whole document ("natural reading"), retrieval via a vocabulary index, the ability to print text, and the creation of statistics on a document, including the number of times a given word or set of words appears in a document.

The Academia Sinica Computer Center began this project in 1984 with a trial data base of the economic chapters of the first eight dynastic histories. The East Asia Library of the University of Washington (EALUW) participated in this pilot project. In 1986, EALUW and CCAS signed an agreement to initiate a joint project to 1) develop a prototype of a Chinese full text processing system, and 2) design an integrated library system. In 1989 a further subagreement was signed by both parties, establishing a cooperative project for the building and use a Chinese dynastic histories data base to be acquired by and maintained in EALUW. The equipment was set up in December 1989 and the Histories tapes were delivered in three shipments between December 1989 and September 1990. The system has been operational since January 1990.

Because of the complexity of this data base, from both its architectural and nonalphabetic-language points of view, it was necessary for CCAS to provide consultation and support for the operation of the system. In September 1990, Mr. Ding Zy-kaan of the CCAS led a team of five people to conduct a workshop at EALUW. This was to provide interested parties with both technical and operational guidance, and data base management techniques.

Presently, the data base system at the University of Washington is the only one in existence outside of Taiwan. The Harvard-Yenching Library is planning to have this system installed by the end of this year. The National Central Library, Taipei, has
purchased the software but has not yet purchased the hardware. Other institutions considering this system are the Institute of Chinese Studies, University of Heidelberg; Tokyo University; Kyoto University; and various universities in the People's Republic of China.

The hardware at EALUW consists of one host machine (AT&T 3B2/310), one system printer (Epson 24-pin printer), one modem (Intel 9600-baud EX), and two computers (Acer). There are also two remote workstations, one located in the Department of Asian Languages and Literatures and the other at the Jackson School of International Studies. Each has an Acer computer with a modem and have dial-in lines to the host station. The Twenty-Five Dynastic Histories Full Text Retrieval system was loaded into EALUW's host machine. This machine also contains supporting software that, for example, allows the display and manipulation of Chinese characters. This system was established with grant funds from the Jackson Foundation, the Charles Fritz Endowment in Chinese Studies of the UW Jackson School of International Studies, and the UW Libraries.

Technical experts from CCAS came to EALUW to set up the system and then conduct a workshop. They also provide technical assistance and improvements to the system if needed. The UW Academic Computing Center is currently working on a way to include the Chinese character-oriented data base in the campus network, as well as making off-campus network connections. The UW Libraries Systems Librarian is managing both the hardware and the software for the Chinese data base, including daily troubleshooting operations.

The EALUW staff provides operational instruction for library users, conducts searches, and sends out search results to off-campus users. The Library staff has also prepared a basic user manual, other necessary search tools, and conducts group instruction for users. CCAS has provided a manual in Chinese and a draft manual in English. Both manuals are seventy to eighty pages long, but lack indexes.

No fees will be charged until 1992. User fee schedules will then be arranged between IHPAS and EALUW. The dial-up remote access is possible for those with the necessary hardware. As of now, the two remote workstations on campus and a station at UC San Diego have the necessary hardware setup. Access through the network connection is still waiting to be developed.

The Histories Data Base provides a full text retrieval capability which is in contrast to other data bases, such as Dialog and ERIC which display only abstract, selective text using keywords. The user may choose to read text page by page on screen—like reading a book—as well as retrieve any words or terms (not keywords) from the full text data base. Boolean searches ("exclusive character sets") are possible by using "and," "or," "but," and "and not" and can actually improve the search results. Users do not need any special computer knowledge to operate this data base; it is an integrated, unified system. The search method always remains the same which makes it very easy and convenient for users to conduct free text searches.

So far, less than thirty people have conducted free text searches on the data base since its inception in January 1990. This figure is lower than we had hoped for, yet the number of times used (and the duration of each use) add up to a figure that could easily represent more than one hundred times.

Publicity necessarily plays a very important role in promoting usage of the data base. We made efforts to publicize this data base through announcements and news re-
leases to the Association for Asian Studies Newsletter, the Shih chieh jih pao (World Journal), Hai wai hsüeh jen (Overseas Chinese Scholarly Journal), Inside Wau, U.W. Libraries, etc. We gave data base demonstrations to seventy people and introduced the system at meetings and workshops to two hundred people. We promoted use of the data base with travel grants from the Jackson School of International Studies; so far, eight off-campus scholars supported by these grants have come to the University of Washington to work on the data base. We gave two presentations dealing with the data base at the last Association for Asian Studies Committee on East Asian Libraries annual meeting and at the Asian Studies on the Pacific Coast annual meeting.

It has been over a year since EALUW began providing use of the Twenty-Five Dynastic Histories Full Text Retrieval Data Base system. In March 1991 a user's questionnaire provided by Library School students at the National Taiwan University was sent out to users of this data base. Fifteen out of twenty-five recipients of the questionnaire responded to it. A summary of the results follows:

1) Education/status
   Master's program students 4
   Ph.D. program students 6
   Faculty members 3
   Librarians 2

2) Disciplines
   History 6
   Language and literature 7
   Library Science 2

3) Reason for use (may select more than one)
   Research-related 12
   Work-related 2
   Individual interest 3

4) Number of times used in the past year
   Once 2
   Twice 1
   Three times 6
   Four times 2
   Five or more times 4

5) How did you learn about this data base?
   Self 3
   Library staff 10
   Other persons 2

6) Opinions about the system

<table>
<thead>
<tr>
<th>Excellent</th>
<th>Satisfactory</th>
<th>Unsatisfactory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed of search</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Convenience of search</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Completeness of search functions</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Clarity of screen</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Satisfaction with search results</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Satisfaction with the whole system</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

7) Suggestions
   - Needs a Wade-Giles or Pinyin romanization input capability.
   - Page references should be to the sentence containing the targeted term rather than to the paragraph.
Full text retrieval systems for English-language materials have been in existence for some time now, and their users have been able to benefit greatly from them. Chinese characters, on the other hand, could not until recently be stored and retrieved via computer. This fact, together with the lack of indexes for classical Chinese materials, makes the study of old Chinese texts very difficult. There are in existence many manual research tools to supplement the study of these texts but these methods of study are very time-consuming and therefore often lead to incomplete or even inaccurate results. Some researchers have created their own indexes for these ancient texts, but these indexes are far from complete and do not include such words as grammatical particles.

This data base is a free text retrieval system, so any word is retrievable and is retrieved in a very short time. Along with the ability to compile statistics on any given document, these are the features of this system that set it apart from existing manual tools of study. The time saved by researchers is immeasurable. This data base system allows kinds of research never before possible. A linguist could use the statistics feature to methodically study sentence structures, grammar, and word usage. In the future, this system may also be used to store Chinese texts created by Academia Sinica as, for example, *Shih san ching* (The Thirteen Chinese Classics), *Chuang-tzu*, *Kuan-tzu*, and Taiwan gazetteers.

There are many benefits to be derived from this new system. Research results are obtained much more quickly and with greater accuracy than ever before. The system can be expanded to include other texts. It easily pays for itself almost every time it is used.