THE USE OF SOURCE MATERIALS ENHANCED BY MODERN INFORMATION TECHNOLOGY†

Margaret C. Fung*

ABSTRACT

Following a general description of the increasing availability of source materials in Roman languages offered to social scientists and scholars in the humanities through online information retrieval services, this paper discusses recently enhanced services rendered by bibliographic utilities to libraries and research projects on Chinese Studies supported by computer-based technology.

With a succinct account of computer-processing of Chinese vernacular materials in Taipei as background information, the author highlights some of the new resources and systems made available to the scholars of Chinese Studies through recent technological developments. The paper is concluded with suggestions for possible future developments and the role that China scholars may play in this endeavor.

INTRODUCTION

The success of research lies greatly in the effective and efficient use of source materials. The availability, accessibility and effective use of source materials determine the quality and fruitfulness of research endeavors. In the last two decades, with the advent of the information explosion, these factors effecting research have drawn much attention. Numerous efforts have been made to aid research by investigating and experimenting with

* Visiting Professor, Department of Library Science, National Taiwan University.
modern technological products. Equipment, such as photocopying devices, microfilming, and computer/information technologies, have, indeed, made remarkable and revolutionary contributions to research activities.

For decades, people associated the use of technological equipment with fields in basic sciences only. However, the use of such tools has gained increasing popularity in the fields of social sciences and the humanities since the 1970's as indicated by McGee and Trees:¹

The 1970's and early 1980's will probably be known to information specialists as the age of the machine-readable, numeric, social sciences data file. Public interest in social problems and the widespread use of computer based systems to administer, transfer programs and conduct evaluation research has resulted in an exponential increase in the amount of computer-readable data being generated.

In particular, the computer's data compilation, analytic and storage capabilities, statistical packages, and advent of bibliographic databases have greatly enhanced the efficiency of research in the social sciences and the humanities. ERIC, Psychological Abstracts, Management Contents, ABI/INFORM, Social Scisearch, Sociological Abstracts, Social Sciences Citation Index, Comprehensive Dissertation Abstracts. New York Times Information Bank, Economic Abstracts International, MLA Bibliography International, Language and Language Behavior Abstracts, Exceptional Child Education Abstracts, Historical Abstracts, Public Affairs Information Service (PAIS) exemplify some of the available databases through BRS, DIALOG, INFORMATION BANK, and ORBIT.² However, these databases cover mostly materials in Roman languages. They can assist China scholars to locate materials in Roman languages, but they provide no data in vernacular languages. This gap is in need of immediate attention.³

Most of the equipment uses (photocopying, microfilming,
etc.) make no difference as far as language of the research resources is concerned; however, they do make a difference when computer technology is employed, especially in the case of Chinese source materials. After many experiments in the 1970's, finally in the late 1980's, we can enjoy the assistance of computer and computer-related technological developments in the use of Chinese materials.

The paper will limit itself to the discussion of services to aid research in Chinese Studies supported by computer-based technology. It will present a brief historical background on computer-processing of Chinese vernacular language materials, and highlight some of the resources made available to the scholars of Chinese Studies through recent computer and videodisc technology. It will also point out directions for possible future developments, and the role that China scholars may play in this endeavor.

INITIAL EFFORTS IN COMPUTER PROCESSING OF CHINESE SOURCE MATERIALS

Current computer processing of Chinese vernacular materials is still in the stage of I/O devices, which enable people to input and output Chinese materials in vernacular form. The computer-assisted use of Chinese source materials was first experimented by library automation projects in the mid 1970's with considerable success.4

The Management Information System in Agriculture Science and Technology (MISAST), ROC Information Database (FICA), Chinese Educational Information Database (CERIS), and the Acquisition, Cataloging and Circulation modules (ACCI) were respectively developed by the Agriculture Science Information Center, Freedom Council Information Center, National Taiwan Normal University, and National Taiwan University. They have proven to be successful prototypes in the 1970s. Unfortunately, some of them are now discontinued.

Subsequent projects on Chinese Character Codes for Infor-
mation Interchange, Chinese MARC formats, the revision of Chinese Cataloging Rules, the adoption of ISSN and ISBN have made substantial contributions to standards essential to operation of Chinese library automation.\textsuperscript{5} As a result, National Central Library's *National Union List of Serials, National Bibliography* and other automated modules have greatly improved both the technical and public services in Taiwan. These projects initiated in Taiwan, as described below, have also benefitted programs in the United States.

**INTERNATIONAL COOPERATION AND THE ENHANCEMENT OF THE USE OF CHINESE MATERIALS**

In the East Asian library community in the United States, several projects using modern technological devices have been in existence for quite some time, such as the East Asian Microfilming Project, the Task Force on Rare Chinese Books, the East Asian Cataloging Subcommittee, and the East Asian Conspectus for Collection under the umbrella of the East Asian Program. The American Council of Learned Societies have also published a report written by the Joint Advisory Committee to the East Asian Library Program.\textsuperscript{6} The report entitled, *Automation, Cooperation and Scholarship: East Asian Libraries in the 1980's* reflects that due attention is given to automation in the East Asian library community in North America. Thus, the need for better support of East Asian research and scholarship is not only a domestic concern in the East Asian countries, but has also become an international concern in the United States and Canada.

Aided by initial and successful efforts achieved in computer technology, and in information interchange codes by the Chinese library, computer, and linguistics communities in Taiwan in the early 80's, two information networks in the United States were able to make breakthroughs enabling them to provide bibliographical control of Chinese vernacular materials and related services to libraries and readers in the United States. In addition,
various recent technological developments such as personal computers, artificial intelligence, CD-ROM and interactive videodisc have greatly enhanced the use of source materials.

Research Libraries Group and Research Libraries Information Network

The above mentioned East Asian Program guides the developments of the Research Libraries Information Network (RLIN). RLIN has supported the input, display, retrieval, storage, and transmission of bibliographic records containing Chinese, Japanese, and Korean (CJK) scripts since 1983.

Currently, RLIN provides approximately one half-million East Asian records to twenty-two participating East Asian libraries in the United States including major East Asian libraries such as Columbia, Cornell, Hoover Institution, Library of Congress, Princeton University, University of Chicago, University of Hawaii, University of Illinois, University of Michigan, and Yale University, etc. These collections represent 75% of the total East Asian Library holdings. Even though RLIN is set-up for cataloging and bibliographic control purposes of libraries, its by product on line union catalog has been extremely helpful to researchers. The database covers books, serials, maps, recordings, scores, visual materials, archival and manuscript materials which can be made available to patrons through its on-line interlibrary loan system. The LC Author Heading authority file and the LSP authority link greatly facilitate convenient searching.

The RLG CJK Thesaurus provides a resource file of 35,000 records, with each record representing an East Asian character. It is an on-line, multi-indexed, multi-lingual dictionary. Each character record shows such information as radical, stroke-count, readings in the three East Asian languages, and how to compose the character. It also shows all variant forms associated with a particular character. It is considered an “East Asian character authority file”, for CJK bibliographic records, and it also serves as an useful dictionary as well.
With PC based equipments, fonts supplied by JHL, and keyboards devised by Transtech International Corporation, the recently developed RLG Multi-script Workstation will help the Network to accomplish its multi-faceted objectives. It features CJK word-processing and CJK interface with other software as well as RLIN functions. Scholars are able to search the RLIN database for relevant materials; download the records to floppies on their PC; massage the data into bibliographies and text; and print the multi-script document off on a printer.

**On-line Computer Library Center**

The largest bibliographic utility in the United States, OCLC, started to explore the possibility of automating East Asian vernacular materials in 1983. A pilot project with the participation of eleven libraries began in May of 1986. Its CJK membership has been increased rapidly with current participation of 65 users including three foreign libraries (National Central Library, University of Alberta Library, and Australia National Library), and other universities and public libraries. With some alterations on their PC/XT based terminal, it offers a variety of input and retrieving methods: Chang-chieh method, Wade Giles and Pin-yin. Wade-Giles and Pin-yin I/O devices have proven to be more user-friendly than other methods. Most scholars in the China field are familiar with both Wade-Giles and Pin-yin. In closer cooperation with National Central Library and other libraries in the Far East, the OCLC CJK database will soon have a large volume of records added to their 110,000 unique vernacular records – almost half of these records are pertaining to Chinese materials). Assisted by US$240,000.00 grant from the Luce Foundation, 100,000 records of Chinese publications with 1911-1949 imprints will be merged into OCLC CJK database. The OCLC CJK program offers services to users with its on-line catalog and on-line interlibrary loan service. It will prove to be even more useful once subject search is made available to users in addition to its original fifteen access points.
The First Emperor of China — Project Emperor-I

Joseph Raben indicated that databases of humanistic materials are beginning to alter the way in which researchers interact with objects of their studies. The videodisc's large storage, integration of multi-type information, high resolution and fast random access capabilities have made these interactions possible.

For example, NEH sponsored the Project Emperor-I which presents and interprets a major historical/archaeological period of China's past. Using the latest videodisc technology, the most spectacular discovery of 7,000 life-size terracotta figures of warriors and horses near Chin Shih-huang's tomb was stored, processed, and made accessible to people who wanted to view and study the artifacts.

Dr. Chen Ching-chih and her team, under the sponsorship of NEH grants, were able to demonstrate the capabilities of the new technologies in supplying users with the improved and more efficient large-scale information processing, storage, access and delivery.

The visual information from two double-sided 12” NTSC CAV videodiscs (each side composed of 108,000 frames of visual images), the visual information is matched with appropriate bilingual narrations, the texts are divided into chapters, very similar to an electronic book. This format allows very speedy retrieval of both the audio and visual information on a thematic topic even though any one of the 108,000 frames of visual images can be retrieved randomly and easily in less than three seconds. Various related factual information stored in a micro-based system when used conjunctively with the videodisc, will enable the user to retrieve one will be able to retrieve specific textual, visual and audio data by means of a micro-based interactive videodisc system.

With the use of DEC's Interactive Video Information System (IVIS), the Project developed a variety of computer-assisted instructional courseware in order to meet the various needs of users who have different levels of knowledge of Chinese art
history and archaeology. In addition to various topical choices of lessons, the main menu includes other options such as:

* References - checking the glossary file or viewing the bibliography file. Items included in the bibliography can also be retrieved and viewed in full-text form;

* Exploration - choosing to view three-dimensional pictures and to browse through the slide collection by going forward, backward, zooming in and out, etc.

* In addition, electronic message is available to patrons. Thus, the interactive videodisc has proved to be an effective wide-ranging educational and experimental tool in the new era of interactive learning and education. It can be used in any of the research projects for the humanities or social sciences.

DEVELOPMENTS OF INFORMATION TECHNOLOGY IN TAIWAN AND ENHANCEMENTS

As with other developed countries, many organizations in Taiwan have also devoted much attention to the improvement of information supply. Libraries and information centers, of course, consider information provisions their raison d'être for existence. The following is a description of recent developments in Taiwan:

Academia Sinica

Academia Sinica, the highest research institution in Taiwan has fully used the new technologies and produced ten authomated modules as follows:

1) Full Text Database for Chinese History

This pioneering project of the Chinese electronic book was launched in July of 1984, and aims at processing Chinese historical texts of the 25 dynasties which contain 60 million Chinese characters. As a prototype, Shih-ho-chih (Treatise on Food and Money) has been implemented on three different systems:
<table>
<thead>
<tr>
<th>Location of Prototypes</th>
<th>Computer Center</th>
<th>East Asia Lib.</th>
<th>Computer Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academia Sinica</td>
<td></td>
<td>Univ. of Wash.</td>
<td>Academia Sinica</td>
</tr>
<tr>
<td>Date of Installation</td>
<td>March 1985</td>
<td>March 1986</td>
<td>September 1987</td>
</tr>
<tr>
<td>Hardware</td>
<td>IBM 5550</td>
<td>Micro-VAX II</td>
<td>AT&amp;T 3B15</td>
</tr>
<tr>
<td></td>
<td>16 bit PO</td>
<td>Dragon 570</td>
<td>Dragon 570</td>
</tr>
<tr>
<td>Operating System</td>
<td>PC-DOS</td>
<td>VMS 4.1</td>
<td>UNIX V;BINARY 1.0</td>
</tr>
<tr>
<td>Query Form</td>
<td>Interactive</td>
<td>Interactive</td>
<td>Interactive</td>
</tr>
<tr>
<td>Access Method</td>
<td>Inverted list</td>
<td>Inverted list</td>
<td>Document-structure</td>
</tr>
<tr>
<td></td>
<td>Full-text scan</td>
<td>Full-text scan</td>
<td></td>
</tr>
<tr>
<td>No. of users</td>
<td>1</td>
<td>2-4</td>
<td>8-48</td>
</tr>
</tbody>
</table>

The databases soon to be completed *Shih Chi, Han-Shu, Hou-Han-Shu, San-Kuo-Chih and Shih-Ho-Chih*. It is expected that the entire database will be completed by June of 1990. Controlled vocabulary (keywords) and free-text searching are currently available. Controlled vocabulary (keywords) are classified into six categories: names, places, time, official titles, references and major events. Six inverted list files were built to include these keywords and the pointers of the corresponding basic text elements.

Materials sought and viewed on the terminal can be copied to the readers’ own files and are to be used along with the readers own notes. Concordances can be generated as well. Statistical analysis for Chinese character/word usage is the most proper application and welcomed feature of this system.

2) Chinese Electronic Dictionary

Based on *Kuo Yu Jih Pao Tze Tien* (*Gwoyeu Ryhbaw Tsyrdan*), the most popular Chinese dictionary in Taiwan, the Computer Center of the Academia Sinica developed a database of Chinese vocabulary and named it “Chinese Electronic Dictionary” (CED). CED is to be used as a machine readable tool for various Chinese information processing applications, especially for the study of Chinese computation linguistics.
This textual database, having 40,000 most-frequently-used Chinese vocabulary words and related information expressed by 1.6 million Chinese characters, serves as an electronic dictionary on line. Each entry consists of Chinese character strings, pronunciation, definition, syntactic categories and annexed attributes, all of which can be used with great ease and efficiency.

With a set of machine-readable files, index files, a number of management programs, and a user interface, CED can provide users with immediate access to the words and related information stored in 3.1 M bytes AT&T computer facilities.

3) Database of Ancient Tombs

Using AT&T 3B/2, UNIX, and Informatics/SDL database management systems, data pertaining to 3040 tombs of Han Dynasty have been processed. The database contains information on each individual tomb's date, structure, location, direction, path, decoration, inscriptions, accompanying buried items, method of burial, and the structure of the coffin. Related statistics on dates, location, and buried items pertaining to any of the above mentioned items can be generated.

4) Household Registers during the Period of 1906-1947

Expert system technology is applied to automate the analysis of household registers for anthropological research. Embedded Structured Query Language in C was applied to model an expert's knowledge of a physical process. The system organizes knowledge into three levels: data, knowledge base, and control. The IBM PC database (with pascal programming language) covers intermediate file (several mega bytes for each district) of 10,338 families from three districts (Chu pei, O Mei, Ta Nei). Records include marriage, family type, adoption, ethnicity, occupation, sibling, age, spouse, etc.

From these records one can reconstruct the marriage and adoption practices for the preceding sixty years. Anthropological research will greatly benefit from the massive amount of data stored and processed with expert. For example, with the manipulation of the household census records, they will be able to find
out information such as:

A. The effect of a person's birth order upon the change of family structure, upon family models, divorce rates and birth rates;

B. The cultural differences and similarities between groups speaking Fukien and Keha dialects in Taiwan;

C. The relationship of mutual assistances among Chinese families;

D. The adaptability of Schachter and Caplow's theory on association in a Chinese society.

5) A Classification of Chinese Verbs for Language Parsing\textsuperscript{15}

The Academia Sinica attempted to establish a better classification for parsing Chinese by adopting current theories of feature-based categorization. The new classification is based on the results of analyzing 16,824 verbs, which found that both syntactic and semantic information have to be used in parsing Chinese sentences. The system provides three types of information about a verb:

A. Its case type with their semantic restrictions and syntactic representations;

B. All possible sentence patterns constructed by the verb and its required arguments;

C. Its syntactic and semantic features, such as dimension, control relations, complement and internal structure of verbal compounds.

6) Land Reports\textsuperscript{16}

The collection of archival materials on Land Reports gathered in Hsinchu area by the Japanese Government during its occupation of Taiwan is the source material used for the study on Taiwan's socio-economic history. This database, consisting of 472 volumes of 250,000 land reports, is essential to the study of Taiwan's economic history, the format of land developments, and the relationship between the land owners and tenants, etc. Pertinent statistics can be generated to verify various hypotheses.

7) Study on Formosan Languages\textsuperscript{17}
With various data collected on the above subject, the following linguistics related research projects have been accomplished:

A. A comparative vocabulary of Formosan languages;
B. Five dialects thesauri (Atayal, Puyuma, Pazeh, Kavalan, and Thao Dialects);
C. Cognates database;
D. Bibliography of Formosan languages;
E. Male and female forms of speech in the Atayalic group;
F. The differences between affirmative and negative speeches of the MAGA dialect.
G. The development and the change of tones as affected by Formosan languages.

8) Masters’ Theses Database¹⁸

This on-line database consists of some 23,111 masters’ theses completed in the Republic of China during the years of 1974 to 1985. For each thesis the following information is included: author, title, adviser(s), degree granting university, college, and name of graduate institute, etc. They can be retrieved with one condition or the combination of multiple conditions.

9) Ming/Ching Dynasties Archival Materials Database¹⁹

Containing approximately 200,000 official reports of Ming and Ching Dynasties, this IBM PC/AT based database supplies information on dates, reigns, emperor names, types, and those responsible for drafting of reports.

10) Academia Sinica Researchers’ Biographical Materials and Works Database²⁰

This database provides biographical materials and publications of 450 researchers at the Academy. The following information is made available on line: name of the researcher, institute and division of his affiliation, publication titles, dates of publications, scope of study, the researcher’s fields of specializations and pertinent keywords.

In addition, the Academy’s note-card system for personal use has been successfully designed and implemented on its history document database. Its capacity of 1K bytes holds thirteen types
of information including content, sequential number, source, subject classification, and keywords. It is, indeed, a welcomed tool for researchers.

National Central Library Bibliographic Databases

Subsequent to the successful formulation of all types of library automation standards as mentioned earlier, the Library Association of China and the National Central Library jointly developed the national plan for library automation in 1980. As part of the inauguration of the program, the National Central Library began to build up an on line union catalog of its own monograph holdings and those of the following libraries: National Taiwan University Library, National Taiwan Normal University Library, National Cheng-chih University Library, Catholic Fu Jen University, and Tamkang University Library.

Also available in the NCL cluster of bibliographic databases are: 21

A. Automated National Bibliography including all Chinese books published in Taiwan since 1981;

B. Periodical and Government Document Indices indexing 800 periodicals published in the Taiwan area and 19 government documents;

C. Union Catalog of Chinese Periodical Literature covering 8000 periodicals held by 170 local libraries in Taiwan;

D. Chinese Rare Books (Shan-pen) On-line Catalog including some 15,000 titles of the printed editions and movable-type editions prior to the Ming Dynasty (1368-1843 A.D.), rare editions after the Ch’ing Dynasty (1644-1911 A.D.), manuscripts and editions revised by famous scholars, and hand-written editions, etc. Since 1983, those rare Chinese books in the categories of classics and history have already been inputted into the database. The remaining three categories are to be inputted into the database in the next two years.

The National Central Library has automated its acquisition,
cataloging and serials control modules, through which the NCL is able to provide bibliographical services to its patrons with its on-line searching capabilities. At the National Central Library, the Center of Chinese Studies is planning an on-line guide covering biographical and bibliographical information on China scholars. When this database is completed, it should prove to be a useful tool in locating subject specialists and pertinent works in the field.

Other Endeavors in Taiwan

Most of the libraries in Taiwan have paid keen attention to the development of automated Chinese library services and have experimented with various projects in the last decade. Some have achieved moderate success; some have discontinued their individual planning and operations in order to wait for the results of united efforts among academic libraries; and some are taking new initiatives in planned networking.

The Tamkang Automated Library Integrated System (TALIS) designed cooperatively with IBM, based on DOBIS, is an integrated system capable of networking on for acquisition, cataloging, circulation, serial control, electronic mail and public on-line catalog. Implemented on IBM4381/M11 system since 1986, it has great potential in networking with domestic and foreign libraries.22

NCTU LIAS developed by National Chiao-tung University since 1983 offers several modules which include an automated monograph system and a serial control system, etc. Each of these two systems has two components: technical services and public service. The public services portion would provide on-line cataloging and retrieving functions whereas the technical service portion aims at acquisition, accession, claiming, and shelving, etc.23

The Science and Technology Information Center (STIC) pioneered library automation with its Union List of Scientific Serials in Roman alphabets in the early 1970's. In addition to what it has accomplished in building databases for scientific/technical periodical literature and personnel data, the Center is
now undertaking the following projects with the purpose of providing up-to-date on-line information retrieval service:

*Establishing the STICNET, a nationwide Sci-Tech information network:
*Integrating domestic R and D information sources;
*Importing selected international databases of great importance.

LEGISIS (Legislative Information System), developed by the Information Center of the Legislative Yuan on VAX 8530 system, offers three useful databases at the present time:

A. Legislators Interpellation System covers more than eleven thousand interpellations made by the legislators since September 1984. All interpellations raised by legislators concerning administration, proposals, and budgets can be retrieved via the name of the legislator, pertinent subjects, government organization's name, keywords, and session number of the interpellations.

B. The full-text database of Acts, Laws, and Codes of the Republic of China: It consists of 560 types of legal documents (12,000 items), and each legal document is described in details regarding its contents and structure. Readers can get access to information with many accessible points such as dates, content summaries, categories, and subjects.

C. LEGISIS THESAURUS on-line database serves as a tool to the above databases, and an on-line dictionary of legal terms.

Two additional databases on newspaper clippings and legal periodical literature index will be available within the next two years.

CONCLUSION

The systems briefly described above, illustrate some of the
computer-assisted research aids and library services which are already available to scholars in the China field. It indicates that bibliographic information is easily made available through bibliographic utilities in a network mode or through individual library services. What is available is mostly pertaining to bibliographic information; people share resources by using an on-line networked bibliographic union catalog.

What will aid the researchers the most in the next decade probably will be full-text databases. As the basis of electronic books, full-text databases will become more useful and effective, if they employ various types of technologies including CD-ROM and videodisc, made available internationally through telecommunication devices. Furthermore, these full-text databases need subject analyses made by subject specialists. Such analyses can best be made if those conducting the research have basic knowledge of computer applications and tools.

More vernacular full-text databases, or databases consisting of vernacular Chinese periodical indexes should be incorporated into existing databases for universal use. Automated union catalog of Chinese Rare Books planned cooperatively by the RLG and the National Central Library should be implemented as soon as possible.26

Standardized database structures, software packages, indexing methodologies, and closely cooperative and coordinated efforts among scholars (users), librarians and computer scientists (the suppliers) are areas to be encouraged. A clearing house of computerized source materials perhaps is an area to be explored by professional associations. Since supply and demand are closely related, no sophisticated tools and resources can be provided unless the researchers make these demands. It is up to the scholars to use these tools, which can be improved with repeated use.

It is the most opportune time for scholars in the China field to look into this important development, and for Professional associations to take appropriate initiatives in coordinating such meaningful efforts.
REFERENCES


毛漢光, “第一期史籍自動化簡介” 計算中心通訊 2:4(民 75 年 2 月 16 日): 13〜14。


13. 中央研究院歷史語言研究所計算中心, “漢代墓葬資料處理系統” 計算中心通訊 2:9 (民 75 年 5 月 1 日): 39〜40。


16. 楊愛憲, “日據時代土地登記書資料的自動化”, 計算中心
通訊 4：13 (民 77 年 7 月 1 日)：99 ～ 100。
17. 李子英，“台灣土著語言資料自動化” 計算中心通訊 3：17 (民 76 年 9 月 1 日)：116 ～ 120。
18. 杭樺，“碩博士論文查詢系統簡介”，計算中心通訊 4：11 (民 77 年 6 月 1 日)：84。
黃佩燕，“博士論文提要系統簡介”，計算中心通訊 3：7 (民 76 年 4 月 1 日)：39 ～ 41。
20. 石玉祥，“本院研究人員著作查詢系統”，計算中心通訊 4：11 (民 77 年 6 月 1 日)：85 ～ 86。
23. 國立交通大學圖書館自動化系統簡介，新竹國立交通大學圖書館，民 77 年。
國家科學委員會科技資料中心，科技性全國資訊網路之規劃及服務功能，民 76 年。
顧敏，“立法索引詞彙簡介”，中國圖書館學會會務通訊 62 (民 77 年 5 月 31 日)：14 ～ 15。