Development of Humanities Computing at Academia Sinica

Drs. Kuan-chung Huang, Tseng-kuei Liu and Jen-der Lee
Institute of History and Philology, Academia Sinica

Contents

* A Chinese version of this article, written by Drs. Kuan-chung Huang and Tseng-kuei Liu appears in *Newsletter for Research in Chinese Studies* (Taipei: Center for Chinese Studies, 66:145-168, May, 1998)

Introduction

The prevalence of information technology and electronic data processing greatly enhances our capability in handling research materials and other non-digital information. In fact, digitized materials have become a major media of research in the humanities. In addition to depending on their individual knowledge and judgment, scholars also rely on computers to locate useful information and improve accuracy. Consequently, the utilization of computers, interpretation of the results generated by electronic media and the construction of related theories have become important tasks of scholars of humanities.

In 1984, the Institute of History and Philology and the Computing Center of Academia Sinica initiated a joint project entitled "Hypertext of the Monographs of Economy of the Twenty-five Dynastic Histories" to promote integration between the institutes of humanities and sciences. Thanks to the efforts of various institutes and the Computing Center, Academia Sinica now possesses the largest and most comprehensive electronic text database for Chinese Studies.

This article introduces our accomplishments in computing technology at Academia Sinica during the last thirteen years and discusses plans for future development. The humanities computing databases at Academia Sinica include the following five categories: First, in addition to an on-line full-text database of over 140,000,000 Chinese characters, a database named "the New Complete Works of the Four Treasures" is currently under construction. Second, databases of research references and tools, such as an "Electronic Dictionary," "Linguistic Corpus of the Academia Sinica," "Sino-Western calendrical converter for two thousand years," and the "Electronic Card System Database". Third, multi-media databases of topical research, such as "Databases of the Research Group for Artifacts and Images," "Taiwan Studies over the Internet," and "Research on Lanyu". Fourth, bibliographies and multi-column databases, which include "Synthetic Research System for Tomb Excavations of the Han Dynasty," "Database of Native Taiwan Languages," "Database of the Household Registration in Taiwan during the Japanese Occupation," "Database of Land Registration Documents in the Zhu-Qian Area in the Qing Dynasty," "Bibliographies of Historical Studies," "Works on Taiwan Archaeology," "Database of Chinese Archaeology," and "the Index of the Grand Secretariat Archives of the Qing Court". Fifth, CD-Roms of rare books and documents, such as "Database of the rare books in Fu Ssu-nien Library," "the Grand Secretariat Archives of the Qing Court," "Database of Economic Documents in the Libraries of the Institute of Modern History and the Institute of Economics" and so on.

The "Database of Full Text Documents"

I. The Development of the "Full Text Chinese Documents Database"

The project of the "Full Text Chinese Documents" began in July 1984 when researchers from both the Institute of History and Philology and the Computing Center at Academia Sinica worked together to key-in the "Monographs on Economy" from the dynastic histories. The original intention was to promote cooperation between the institutes of humanities and sciences and to provide scholars in Academia Sinica with a speedy and powerful search mechanism by digitizing classical texts. In view of the excellent results, in 1986, the project

was expanded to include the entire twenty-five dynastic histories. In June 1990 the computerization of the full text of the dynastic histories was completed with the exception of the "charts." (Now the charts are all included.) The "Database of the Twenty-five Dynastic Histories" is the first and largest segment of the Full Text Project, and has been well received by scholars around the world.

After the completion of the computerization of the dynastic histories, the project was expanded to include many other documents according to research needs of various institutes. These were the "Thirteen Classics," Buddhist texts, traditional Chinese medical texts, political documents and linguistic corpus (by the Institute of History and Philology), Series of Taiwanese Documents (by the Preparatory Office of the Institute of Taiwan History), and many other research related documents. By 1998, this project had completed approximately 140,000,000 Chinese characters (see Table I). There are other databases involving 200,000,000 Chinese characters currently under construction (see Table II). On completion, the Full Text Database will be the largest database of computerized Chinese texts and one of the most important resources for Chinese Studies in the world.

Careful attention has been paid to the selection of the versions of books for key-in, and careful proof-reading on documents lowers the mistake rate to 1/1000. It is the largest database of its kind with 300,000,000 characters at hand and 10,000,000 more characters added each year. Certainly, compared to traditional paper documents, digitized full text documents are more accessible, less expensive and easier to store.

Scholars who have used the Full Text agree that the most important aspects of the database are its speed and comprehensiveness. Traditionally when historians collected material for research, they would carefully read through often lengthy documents, write down related sentences/paragraphs, and then copy these into their drafts again when they wrote up their articles. This process takes a tremendous amount of time and runs the risk of accidentally missing an important piece of evidence. Now with the Full Text, scholars are able to search through huge amounts of documents in several seconds without fear of overlooking an important entry. Take, for example, the system set up on UNIX in the Institute of History and Philology. It takes only one or two seconds to search one character/term throughout the twenty-five dynastic histories, a collection containing some 40,000,000 characters in total. Users can read the documents on screen and print/save useful material after primary selection. The newly established WWW search system is even more accessible than the DOS version.

The "Full Text Chinese Documents Database" not only speeds up search but also improves accuracy for working scholars. As a result, since its establishment, it has been purchased by universities and research institutes all over the world and assisted scholars in countless research projects.

II. The Construction and Application of the Databases

1. The Construction

To help various institutes in Academia Sinica develop databases, the Computing Center has created a software which includes a program for construction and a program for searches. The former constructs marked up digitized documents to databases and the latter locates information from databases according to the search conditions of the users. The whole process of constructing the databases includes key-in, proof-reading, making characters,

markup and organization.

1) Key-in: Two groups of people type up documents making two copies of the electronic version.

2) First proof-reading: The two copies on computer with a proof-reading program to locate discrepancies. Although comparing two copies by computer doubles the expense of key-in, it nevertheless costs less time and expense than manual proof-reading.

3) Second and third proof-reading and markup: Professional assistants check and correct the electronic versions of the documents manually with original books, and then mark up the electronic documents. To mark up is to show the organization of the documents, such as volumes, chapters, sections and paragraphs as well as editorial structure, so that the construction program can establish files of texts and index the databases.

4) Making characters: Words in classical texts that cannot be found in Chinese computer systems need to be recognized, collected and created on computer. By now we have made over 4,555 Chinese characters for the Databases of Full Text Chinese Documents.

The third step in the above process is the most expensive and time-consuming. Classical texts often do not use punctuation. Professional assistants or even scholars are often needed in separating the chapters and sections in certain texts. In the early stage of the construction, Academia Sinica had to spend NT$1 for each character from key-in to markup; now the expense has been cut down to NT$0.5 since experience has helped improve efficiency.

2. Application and Accessibility

The construction of the databases is almost entirely supported by the regular financial budget of the various institutes involved. The results so far have contributed to the long-term cooperation and improvement of the Institute of History and Philology and the Computing Center. In the past, only institutes that have purchased the databases had the right to use them so that the financial support of the databases could be guaranteed and the pay for use principle could be maintained. Now with the development of the internet system and our intention to contribute to society, we think it is the proper time to open the databases to academic and educational institutes. The regulations of usage and payment of the databases were revised in March 1997 after a year-long experiment:

1) Usage: Users can connect to the database and search through World-wide web.

2) Free Usage: 30 entries for Full Text Chinese Documents and 2000 entries for the Corpus. We will increase the free usage entries each year.

3) Paid Usage: Established databases, except those with IP problems, are all open to institutions in Taiwan with annual charges. All money received will be handed to the national treasury.

4) Free Usage for Cooperation Institutions: Institutions that have cooperated with Academia Sinica in developing the databases are granted a certain number of the databases and free usage during the time of cooperation.

Moreover, the Full Text Chinese Documents Coordination Committee of Academia Sinica has decided to develop the "Databases of Humanities for Teachers and Students" and opens it free of charge to teachers and students in all levels of schools. The purposes are threefold: to promote research and education of humanities in Taiwan, to fully share and apply resources and to increase the cooperation between humanities and sciences. The contents of these databases are selected and edited from the existing databases by the Editorial Committee of the Humanities Databases (for teachers and students), which includes five people from Academia Sinica and five teachers from elementary, junior high and high schools. These databases contain about 40% of all the databases and cover a wide range of resources; the process of opening the databases to the public is still in a preliminary stage, but has at least begun.

III. Future Development: the New Complete Works of the Four Treasures

1. Background

The improvement of both quality and quantity of the databases is required to maintain the position of our databases in academic research and educational promotion. The Full Text Chinese Documents Coordination Committee of Academia Sinica, set up in October 1996, seeks this improvement in two ways. First, we have established the Work Station for Chinese Studies, headed by Professor Hsieh Ching-chun, to improve software design and to speed up the search function. Second, we plan to expand the contents and research fields of the databases through cooperation with other academic institutions. "The New Complete Works of the Four Treasures" is designed under these considerations.

Some publication companies and academic institutions in Hong Kong and Mainland China have initiated projects to digitize "the Complete Works of the Four Treasures" within one and a half years. Due to commercial and market concerns, these projects often sacrifice academic quality for production deadlines. In order to avoid these pitfalls and to provide the scholarly world a better product, Academia Sinica has decided to construct "the New Complete Works of the Four Treasures."

2. The Characteristics of the original "Complete Works of the Four Treasures"

As the most voluminous collection of traditional Chinese texts, the "Complete Works of the Four Treasures," compiled in the 18th century, rightly attracts the attention of publication companies when they publish digitized books. However, after several meetings of the Full Text Chinese Documents Coordination Committee of Academia Sinica, we have decided not to follow in the foot steps of the commercial publication companies because the original collection is flawed in many ways.

First, the contents of the collection are not comprehensive and the quality of the texts selected are not good. Under the authoritative rule of the 18th century Manchurian emperor Qianlong, officials who compiled the collection discarded and burnt quite a number of texts which contain the histories of the Manchurian invasion in the 17th century. Books that were perceived as dangerous to imperial rule were also rewritten or partially rejected. Therefore, the quality of versions selected in the collection is often questionable.

Second, the collection mainly contains texts either written by or about Confucian scholar officials, while religious documents, technological texts and materials of popular culture are

often left out. Third, since the collection was compiled in the 18th century, texts after that period were not included.

In view of the above problems, the Committee, after several meetings, decided to construct the "New Complete Works of the Four Treasures".

3. The Scope of the "New Complete Works of the Four Treasures"

The "New Complete Works of the Four Treasures" is tentatively designed to include the following five areas so that it can integrate current research topics with future research needs of the institutes of humanities in Academia Sinica and at the same time surpass the original "Complete Works of the Four Treasures" in scope and in volume. These five areas are, first, the original "Complete Works of the Four Treasures," including published texts (over 79,000 volumes, 17,000,000 pages), texts with only titles (93,551 volumes, extant now over 60,000 volumes), and appendix texts (over 30,000 categories). Second, religious texts, including Buddhist and Taoist documents and documents of popular religions. Third, literature, including the *Chu tz'u*, compilations of literature, critiques on poetry, dramas, novels and lyrics. Fourth, recipes and techniques, such as texts on medicine, health, geomancy, divination, fortune-telling, and so on. Fifth, historical documents on Taiwan, including archives, local gazetteers, contracts, literature, biographies, statistics, Japanese materials, linguistic materials and inscriptions.

4. Guidelines and work to be completed this year

The above-mentioned materials include 200 billion Chinese characters, about three times the volume of the original "Complete Works of the Four Treasures." It requires human and financial resources that go far beyond what Academia Sinica can provide. In order to execute this project, the Coordination Committee hopes 1) to decide priorities in constructing the databases according to current research needs of humanities institutes of Academia Sinica; 2) to lower the expense and to enhance efficiency through cooperation with institutes outside Academia Sinica; 3) to integrate high quality databases (such as the "Database of the Complete Tang Poetry") into our databases through purchase; 4) to section and to punctuate classical texts to accelerate the search function.

The construction of the "New Complete Works of the Four Treasures" is a long-term project. The Coordination Committee has decided to use this year as the year of planning, proposing long-term guidelines through various experiments. Besides the Work Station for Chinese Studies that will improve the software system, this project will include the following tasks for this year: 1) to invite scholars on version-comparison and cataloguing to provide information; 2) to invite specialists of database construction to provide advise; 3) to purchase good quality databases, and 4) to hold conferences on full text documents to exchange experiences. We are confident that this project can succeed if we obtain financial and human resources supported by institutions all over the world.

Other Databases

Other than the "Databases of the Full Text Chinese Documents," the institutes of humanities in Academia Sinica have also developed 37 databases (categorized into 4 groups, see Table III) to manage archaeological materials, artifacts and images, topical documents, research

references and tools, as well as multi-media CD-Roms.

1. Research Tools, Dictionaries and Linguistic Corpus

One category of the databases is reference tools (see Table III, no.1-4). The "Electronic Card System Database," which imitated a traditional card system in research, was developed in 1986 to help scholars categorize, print out and locate information before writing the articles. It is now out of use. Later on, an "Electronic Dictionary " (now renamed as "Chinese Knowledge Information Processing") was developed in 1988 to serve the needs of linguistic studies. After ten years' construction, it now contains almost 100,000 words and is accessible for search, statistical analysis and print-out.

"Sinica Corpus," recently put on the Web, is the first Chinese corpus that has comprehensive markups for the parts of speech. It contains 5,000,000 modern Chinese linguistic sentences. The markups for the parts of speech, just like the words, can be searched, screened, and calculated for collocation. The building of the "Sinica Corpus" began in 1990 by the CKIP group of Academia Sinica, headed by Drs. Keh-jiann Chen (Institute of Informative Sciences) and Chu-ren Huang (Institute of History and Philology) successively. The project receives financial supports from the Chiang Ching-kuo Foundation, Academia Sinica, and the National Council of Sciences. The first version was completed in July 1995 and put on telnet for use in December the same year. Since November 1996 the WWW version is open for use. (See http://rocling.iis.sinica.edu.tw for more information.)

In the field of traditional Chinese languages, there is the "Database of Traditional Chinese Corpus" (see Table I). In historical studies, the "Sino-Western calendrical converter for two thousand years," developed by Prof. Yeh-chian Wang (Institute of Economics), is a very useful tool to locate and convert calendrical information between China and the West.

2. Multi-media Databases of Topical Studies

This category of databases (Table III, no. 5-8) are constructed around research subjects and they often include many small databases with different features. They are all constructed on the WWW.

The "Databases of the Research Group for Artifacts and Images" was created by the "Research Group for Artifacts and Images" of the Institute of History and Philology. It contains 6 databases: "Database of Bamboo Slips, Silk Books and Bronze Inscriptions-- Full Text," "Database of Bamboo Slips, Silk Books and Bronze Inscriptions-- Bibliography," "Bibliography of Research on Han Reliefs," "Images of Han Bamboo Slips from Ju-yan," "Reliefs from the Wu Family Shrine of Shandong," and "Han Reliefs from the Dong Village in Anqiu." The former three are text materials and the latter three include also images. The "Database of Bamboo Slips, Silk Books and Bronze Inscriptions" contains 44 kinds of materials, 3,401,684 characters, ranging from full texts of bamboo slips and silk books, stone and bronze inscriptions and seal engraving of the pre-Qin and Han periods as well as related bibliographies. This kind of excavated materials are fragmentary and scattered. They do not have logical lay-out of chapters and pages as classical texts do. Such information would be difficult to read and use if not for the digitized full text retrieval system. Now that the databases combine texts and images, researchers can easily locate both texts and related images. These databases have been constructed since 1989 and are available on the WWW since 1997.

"Taiwan Studies over the Internet," headed by Prof. Fu-san Huang of the Institute of Taiwan History, was constructed in 1997 by scholars from the Institutes of Taiwan History, History and Philology, Ethnology, and Sociology. It is developed on the WWW to integrate all kinds of resources on Taiwan Studies, including pictorial and textual electronic books, bibliographies and chronicle charts for search, dictionaries, statistics, materials from field work, and introductions to historical documents and research papers. It incorporates research tools, sources and publications, and is therefore very useful for researchers.

Two databases, "Research on Lanyu" and "Research on the Northeast Coast of Taiwan," are experiments of the Digital Library, a cooperative project between IBM and Academia Sinica. The former is on line for use already, and the latter is still under construction. "Research on Lanyu," presented via textual, and pictorial media, combines materials of ethnography, linguistics, geography, zoology and botany.

3. Multi-columned Databases of Index and Bibliographies

Databases of index and bibliographies are basic research tools; they are often developed first and applied most (Table III, no.9-32). This kind of database is usually multi-columned. Materials are put into columns for easy search. For instance, in the "Synthetic Research System for Tomb Excavations of the Han Dynasty" materials are categorized into 400 items. Users can either locate a certain item, such as a piece of burial vessel, or do statistical analysis. Earlier Academia Sinica had several databases of this kind, including "Synthetic Research System for Tomb Excavations of the Han Dynasty," "Database of Native Taiwan Languages," "Synthetic Research System for Household Registration" by the Institute of Ethnology, "Database of Land Registration Documents" by the Sun Yat-sen Institute for Social Sciences and Philosophy, and "Database of Master Theses and Doctoral Dissertations in Taiwan". They were all built on UNIX using INFORMIX, and none is in use now except for the "Synthetic Research System for Tomb Excavations of the Han Dynasty" and the "Synthetic Research System for Household Registration".

Later on, the Computing Center of Academia Sinica developed DORE (Document Retrieval System, improved later on as DORE II, used also on the Web) so that the columns can be more adaptable to accommodate not only indexes and bibliographies but also paper abstracts or even full text. Several databases are under DORE (see Table III, no. 14-22), such as "Database of Chinese Archaeology", and are now for internal use only. The "Index of the Grand Secretariat Archives of the Qing Court" and "Bibliographies of Historical Studies", however, are now converted to TTS system, and not maintained by DORE anymore.

TTS is a full text retrieval system (with both Web and PC versions) developed by Transmission Book & Microforms Co. LTD. It handles voluminous materials with high speed and cross-column search; it can also connect located materials with other texts and images via hyperlink. Lately many index and bibliographies are constructed under this system (Table III, no. 23-31). They are all on line for use and their quantity is expanding continuously.

4. CD-Roms of Archives and Rare Books

Many valuable materials are kept in the collections of various institutes of humanities in Academia Sinica. For instance, the Institute of History and Philology has 311,914 documents from the Grand Secretariat Archives of the Qing court, over 25,000 rubbings, and 42,000

volumes of rare books. The Institute of Modern History has governmental archives, 18,898 cases on economics (1903-1980) and 3,911 cases on foreign affairs (1880-1928), 34 cases (593 volumes) on the Feb. 28 Incident. The Institute of Taiwan History keeps 202 sections, 9,023 pieces of traditional documents of Taiwan, including land contracts, genealogies, accounts, books, stamp molds, song books, household registration and interview records. The Institute of Ethnology collects 4,335 examples of various sorts of contracts, both original and copies. All these documents are scanned and made in CD-Roms either because their size prevents full text key-in or because their forms and writing styles invite research. Making CD-Roms not only better preserves the original information but also prevents direct handling of the documents which may cause damages.

The Fu Ssu-nien Library of the Institute of History and Philology started to construct the "Database of the precious and rare books" as early as 1988. The library has by now finished over 9,000 volumes, applying the above-mentioned TTS to search and to connect images. The Institute of History and Philology also applied TTS to the "Grand Secretariat Archives of the Qing Court," first entering names, official titles, times, events, and document categories for search in early 1995, and integrating the database with imagery CD-Roms in 1996. The project is still in progress, with about 20,000 documents handled each year.

Several databases of archives are now under construction, including "Archives of the Economic Peace Committee of the Administrative Yuan," "Archives of the Financial Support of the U.S.A." (Institute of Modern History), "Archives of the Production Management Committee" (Institute of Modern History, Institute of Economics and Archival Committee of Provincial Government of Taiwan), "Archives of Patent Bureau of Taiwan under Japanese Occupation," " Government Archives of Taiwan under Japanese Occupation," "Li Kuo-ting' s Documents," and "Old Contracts collected in Academia Sinica" (Institutes of Modern History, Economics, Taiwan History, Sun Yat-sen Institute for Social Sciences and Philosophy, and Archival Committee of the Provincial Government of Taiwan). Some of the archives are now being scanned and some have already been made into CD-Roms. All these projects are supported by the Archives Coordination Committee of Academia Sinica, established in 1996.

5. Future Development

Although not as famous as the Full Text Chinese Documents, the humanities databases mentioned above are mostly in good condition and available for use now. Scholars have also utilized these databases to write papers. For instance, 110 articles, 16 technical reports and 7 volumes of collected papers were published during the construction of the Sinica Corpus.

Among the four categories discussed above, databases of research tools and references are so far least developed. The Electronic Card System is not in use anymore, but similar tools need to be built for research. Hopefully we can develop an interface to format located entries into cards so that scholars can annotate, categorize, organize, and print out to speed up the writing of articles. Moreover, humanities scholars need more electronic dictionaries of, for instance, characters on the oracle bones, bronze inscriptions and so on. Bibliographies are basic references for research and will be constructed continuously. Bulky materials such as oracle bones and archaeological excavations collected by the Institute of History and Philology can also be made into CD-Roms to enhance search and study.

New computer technology has to be developed to accommodate new materials and databases. One of the difficulties scholars have constantly encountered is making characters. In addition

to characters written in different styles, there are also decipherable and indecipherable characters that need to be formulated on computer. These characters often appear on ancient remnants such as oracle bones, bronze and stone inscriptions, bamboo slips and wooden strips. For instance, over 4,000 characters were created during the construction of the database of the Han Bamboo Slips. With the help of Prof. Ching-chun Hsieh of the Institute of Information Sciences, hopefully this problem can be better solved in the future. Furthermore, oracle bones and bamboo slips are often broken and fragmented. It will be invaluable if there were a program to digitize their forms and writing styles, so that scholars could restore the fragmentary pieces into their original structures through computer.

Storage and transmission are two major issues for CD-Roms application. Since the volume is huge, it often takes a long time to transmit the information on computer. Backup tapes are used recently, on an experimental level, to improve the situation. Another way to solve the problem is to use multi-layer rewritable CD-Roms. Romanian scientists have recently developed a new kind of multi-layer CD-Rom; the capacity of which is 10,000 times that of the current ones. Hopefully this new technology can better solve the problem of storage. As for transmission, the most difficult condition happens in voluminous databases of images, such as multi-page images of the rare books. These databases are difficult to transfer on line, and the problem will have to wait for new technologies on image-digitization and internet transmission systems to be solved.

There are four other issues that need to be considered in the future. First, to take full advantage of the expeditious development of the WWW, humanities databases should also enter the multi-media era. Second, since precious and rare books are traditionally under strict rules of usage, such as limitation on photo-copying, a new set of regulations will have to be set up when these databases are put on line for use. So far, institutes that have participated in the construction have not come up with any consensus yet. Third, coordinate organizations for databases should be established to integrate supporting resources, especially those extensive databases with many institutes and enormous financial and human resources involved. Academia Sinica has by now founded the Full Text Chinese Documents Coordination Committee and the Archives Coordination Committee. However, new committees need to be established to set up a unified format and markup programs for CD-Rom databases so that databases developed by both Academia Sinica and other institutes can be communicable and combined for usage, and that programmers can develop programs under unified format. Fourth, the construction of databases requires huge amounts of financial and human resources that go beyond what a single institute can afford. It is essential for Academia Sinica to cooperate with other institutes and to obtain governmental support in building these huge databases. Moreover, coordination and integration of different institutional resources can avoid redundant construction and save resources.

Conclusion

Upon constructing the Gutenburg Project on line in 1971, Michael Hart pointed out that the most substantial contribution of computers was not computation; it was rather the power to store and to search tremendous information that traditionally were stored in the library. The project was named Gutenburg to signified another media revolution after the birth of the printing technology. (See http://www.promo.net.pg/history.html for more information.) More than 20 years have passed, and the application of computers in storing and searching documents, such as texts, archives and artifacts, has proven itself essentially important and

helpful.

The humanities computing at Academia Sinica has provided an excellent example of what can be done. Overall the humanities computing has been well developed, with many kinds of databases and good organization and management; the future of cooperation with other institutions is also foreseeable. However, the most important issue is not the construction but the application of databases. Interaction between researchers and the information, between different researchers on line, and between researchers and the society have to be explored. In other words, as in the case of most new research tools, humanities computing is most helpful when scholars not only use the technology to replace traditional methods of research, but also explore ways of exploiting the technology for new and innovative forms of research.

Table I The "Database of Full Text Documents" of Academia Sinica
2/16/1998

Database	Number of characters	Institute	Coordinator
The Database of Full-text Chinese Document: 1.Database of the Twenty- five Dynastic Histories 2. Philosophies 3. The Thirteen Classics 4. Eighteen classical texts 5. Thirty-four classical texts 6. Buddhist texts	39,969,533 7,267,407 8,600,316 8,049,602 12,264,715 10,118,213	Institute of History and Philology and the Computing Center at Academia Sinica	Dr. Fu-Shih Lin
The Database of Traditional Chinese Corpus	31,736,564	CKIP group of Academia Sinica	Drs. Chu-Ren Huang & Pei-Chuan Wei
Local Gazetteer of Taiwan	7,537,840	Institute of Taiwan History	Dr. Su-Chuan Chan
Documents of Taiwan	7,100,885	Institute of Taiwan History	Dr. Su-Chuan Chan
Diaries of Modern Chinese History	2,086,513	Institute of Modern History	De. Kuo-Tai Hu
Wen Xin Diao Long	1,700,011	Institute of Information Science	De. Ching-Chun Hsieh
Three Discourses on Buddhist Texts	104,257	Institute of Information Science	Dr. Ching-Chun Hsieh
Yao Ji-Heng Collection	951,560	Institute of Literature and Philosophy	Dr. Chiu-Hua Chiang
New Qing History-Biographies of the Emperors	878,629	National History Museum	Mr. Chung-sheng Chu
"Yue-Fu Poem" Collection	633,151	Department of Chinese Literature, National Taiwan Normal University	Prof. Hsu-Sheng Chi

Table II The Databases of Full-text Documents of Academia Sinica
Currently under Construction
2/16/1998

Database	Numder of Characters	Institute	Coordinator

Database of Full-text Document II-	181,785,000	Database of Full-text Document Project of the Institute of History and Philology	Dr. Fu-Shih Lin
Serials of Taiwan Archives	12,700,000	Institute of Taiwan History	Dr. Su-Chuan Chan
Taoist Documents and Plus: 1.Taoist Documents 2.Liu Zong-Zhou Collection 3.Quan Weng-da Collection	 1,891,000 1,100,000 1,000,000	Institute of Literature and Philosophy	Drs. Fong-Mao Lee & Tsai-Chun Chung
Database of Traditional Chinese Corpus II	8,120,700	Institute of History and Philology, Institute of Linguistics, Institute of Information Science	Dr. Fu-Shih Lin Dr. Chu-Ren Huang and Dr. Keh-Jian Chen
The Collection of the Qing Dynasty	579,000	Institute of Modern History	Dr. Kuo-Tai Hu
Buo Re Jing Lun	1,500,000	Institute of Information Science	Dr. Ching-Chun Hsieh

Table III Other Databases at Academia Sinica

	Database	Number of Entries	Institute	coordinator	system	note
1.	Electronic Card System Database		Computing Center			
2.	Electronic Dictionary	over 40,000	the CKIP group, Institute of Information Science			
3.	Sinica Corpus	350,000,000	Institutes of Information Sciences, History and Philology, and Linguistics	Drs. Chu-Ren Huang & Keh-Jian Chen	Corpus	
4.	Sino-Western calendrical converter for two thousand years		Computing Center			
5.	Databases of Artifacts and Images		Institute of History and Philology	Dr. I-Tien Hsing	www	
6.	Taiwan Studies over the Internet		Institute of Hiistory and Philology,	Dr. Fu-San Huang		

			Institute of Taiwan History			
7.	Research on Lanyu		Institutes of Ethnography, Zoology, Linguistics, Botany and Geography		Multi-media CD-Roms	
8.	Research on Northeast Coast of Taiwan		Computing Center and Institute of Zoology		multi-media CD-Roms	under cons-tructio n
9.	Synthetic Research System for Tomb Excavations of the Han Dynasty	3040 entries, 400 items	Institute of History and Philology	Dr. Mu-Chou Poo	INFOR MIX	
10.	Database of Native Taiwan Language		Institute of History and Philology	Dr. Jen-Kuei Lee	INFOR MIX	
11.	Synthetics Research System for Household Registration	Household: 18,000; Individual: 20,000; Events: 20,000	Institute of Ethnography	Dr. Ying-Chang Chuang	INFOR MIX	
11.	Database of Land Registration Documents	about 50,000	Institute of Social Sciences and Philosophy	Dr. Yen-Hsien Chang	INFOR MIX	
12.	Database of Master Theses and Doctoral Dissertations in Taiwan	23,111 (1974-1986)	Computing Center		INFOR MIX	
14.	Index of the Grand Secretariat Archives of Qing Court	49, 720	Institute of History and Philology	Dr. Cheng-yun Liu	DORE II	Interna l use
15.	Bibliographies of Historical Studies	36,852	Institute of History and Philology	Dr. Cheng-yun Liu	DORE II	Interna l use
16.	Database of Chinese Archaeology	12,349	Institute of History and Philology	Dr. Chuan-Ying Yen	DORE II	Interna l use
17.	Database of Economic Documents of	6,707	Institute of Modern History	Dr. Tsui-Hua Yang	DORE II	Interna l use

	Taiwan					
18.	Database of Foreign Affairs Archives	163,232	Institute of Modern History	Dr. Tsui-Hua Yang	DORE II	Internal use
19.	Database of Chinese and Foreign Maps		Institute of Modern History	Dr. Tsui-Hua Yang	DORE II	under cons-tructio n
20.	Database of Taiwan Economic Development	4,173	Institute of Economics		DORE II	
21.	Database of Events of Taiwan under Japanese Occupation	1,526	Institute of Economics		DORE II	
22.	List of "Database of Full Text Documents"	1,318	Computing Center		DORE II	
23.	List of publication of the Institute of Ethnography	666	Institute of Ethnography		TTS www	
24.	Publications of research fellows in the Institute of Ethnography	2,227	Institute of Ethnography		TTS www	
25.	The Abstracts of Socialogy Papers in Taiwan	2,105	Institute of Ethnography		TTS www	
26.	Abstracts of Sociology Publications in Taiwan	3,288	Institute of Ethnography	Drs. Kuo-Shu Yang, An-Bang Yu and Kang-Hui Yeh	TTS www	
27.	Database of Folk Religions Bibliographies		Institute of Ethnography		TTS www	not in use yet
28.	Grand Secretariat Archives of the Qing court	83,887	Institute of History and Philology	Dr. Cheng-Yun Liu	TTS www	
29.	Research Articles on Tang Sung Ming Qing Histories	63,628	Institute of History and Philology	Dr. Cheng-Yun Liu	TTS www	
30.	Publications of Research Fellows of Institute of History and	2,281	Institute of History and Philology	Dr. Cheng-Yun Liu	TTS www	

	Philology					
31.	Index of Archives in Institute of History and Philology(The Fu Ssu-Nien Archives etc.)	3,989	Institute of History and Philology	Dr. Cheng-Yun Liu	TTS www	
32.	Grand Secretariat Archives of the Qing court (images)	83,887 constructed	Institute of History and Philology	Dr. Cheng-Yun Liu	TTS CD-Roms and Novell	internal use
33.	Database of precious and rare books (images)	21,623 constructed	Institute of History and Philology	Ms. Duan-Hsiu wu	TTS CD-Roms, Novell	under cons-tructio n
34.	Rare books in the Institute of Ethnography	4,335	Institute of Ethnography, Archives Coordination Committee of Academia Sinica		TTS, CD-Roms	under cons-tructio n
35.	Government Archives of Taiwan under Japanese Occupation	5,000,000	Institutes of Modern History, Economics, Social Sciences and Philosophy, Computing Center, Archives Coordination Committee of Academia Sinica, and Archives Committee of provincial Government of Taiwan		CD-Roms	under cons-tructio n
36.	Archives of Patent Bureau of Taiwan under Japanese Occupation	4,500,000	Institutes of Modern History, Economics, Taiwan History, Social Sciences and Philosophy, Computing Center, Archives Coordination Committee of Academia Sinica; Archives Committee of		CD-Roms	under cons-tructio n

			provincial Government of Taiwan			
37.	Li Kuo-Ting's Documents		Computing Center, Archives Coordination Committee of Academia Sinica		TTS, CD-Roms	under cons-tructio n