THE DEVELOPMENT AND USAGE OF THE OVERSEAS SINOLOGY DATABASE

Ling Bao

Library of the Chinese Academy of Social Science, Beijing, 100732
Email: Baoling@cass.org.cn

ABSTRACT

The Overseas Sinology Database is composed of three databases: scholar, organization, and journal. The thesis database is regard as separate and is attached to the scholar database. The database information comes from major areas of the world, especially the countries adjacent to China, and updates are done continuously. The Sinology Database is in several different languages and should satisfy the differing needs of data collection and database application. The data quality is strictly controlled during the whole data life cycle, which includes data collection, processing, storage, and accessing. In addition, according to the standards and specifications of the metadata, metadata are created to accompany the data, which satisfies the cooperation among different databases. Finally, besides the function of searching, statistical calculation, and sorting, the database is also used for data mining and knowledge discovery. Through these methods, conclusions about changes in Sinology can be drawn, which will aid us in understanding the world and China in particular.

Keywords: Overseas Sinology Database, Database development, Database usage, Sinology

1 BACKGROUND AND MEANING OF THE DATABASE

At the end of 1970s, the Information Institute of the Chinese Academy of Social Sciences built a ‘Sinology Laboratory’ to do research and publish a journal called *Overseas Studies on China,* with Mr Sun Yue as editor in chief. *Overseas Studies on China* is the earliest journal about Sinology in the Chinese academic community. [1]

With the opening-up and reform policy, China achieved great development and became stronger. Foreign governments, academic communities, and institutions began to take a great interest in China and enhance their Chinese studies. The disciplines that broadened include politics, economics, sociology, history, philosophy, law, ethnics, and education. During this period, Chinese academic communities such as the Chinese Academy of Social Sciences, Peking University, Beijing Foreign Languages University, and the Shanghai Academy of Social Sciences improved the progress of studies on overseas Sinology. Also, governments and diplomatic departments paid more attention to overseas views on Chinese problems. Therefore, overseas Sinology studies have academic and real value [2].

The Overseas Sinology Study Center in the Chinese Academy of Social Sciences established a tenet according to the current need for study. The tenet reports overseas studies on China by international intercommunication, acting as a bridge to connect scholars and institutions from different countries, publish the products of overseas studies and cooperate with foreign institutions doing comparable work. According to this tenet, an important mission is to build an overseas Sinology database [3].

2 THE FEASABILITY OF BUILDING AN OVERSEAS SINOLOGY DATABASE

We considered three principles when building the overseas Sinology database: first, whether this database would have a high academic value and hold a advanced place for a long period; second, whether the database would have a strong document guarantee for the studies of overseas Sinology and fill a gap in China; and third, whether the database would improve social and economic development. According to the above principles, the basic information of the CASS institute meets the requirement of a Characteristic Library Collection, the human resources and material resources that the database required. After the database has been built, it can be used by academia, especially for add-value service. In a word, building the database was feasible.

3 DEVELOPMENT AND CONSTRUCTION OF THE DATABASE

3.1 Composition of the database

The overseas Sinology database is composed of three sub-databases: scholars, organizations, and the journal of overseas Sinology. Beside these three sub-databases, the scholar database describes the works and theses of the scholars whose information has been collected in the scholar sub-database.

3.2 Faculty and administration

Requirements for the faculty who would build the database were research ability, wide knowledge, and a specific specialty. The people who collect and process the data should have the following abilities: knowledge of the information, for example, a familiarity with all kinds of information resources, which they are able to search and access; the ability to judge the authenticity, dependability, and authority of the information resources; and the ability to translate material in foreign languages into Chinese. The ability to use computers to process the data and edit text are also necessary. A professional team, selected according these requirements, was built.

The system administration was based on target-orientation. Material in different languages and from different countries was assigned to individual administrators. They have their own duties and are in their own levels in the organizational pyramid. In this way, initiative, activism, and creativity were inspired so that the goal of quality control could be achieved [4].

3.3 Software

The first step in building a database is to consider the management software. One way to do this is to develop a new system, and the other is to select one that has already been developed. Developing a new system requires technicians who are versed in computer programming and a long period of time in which to complete the task. However, the current situation is short of technicians and time. Thus the feasible method is to select existing software. A library auto-system is not suitable for complicated data structures such as an overseas Sinology database. Without professional training, using library software is a problem. Other management software provides functions such as data uploading, data preprocessing, searching, user management, and material management. The browser/server/database server model is widely accepted in the information industry. Users can build their own service systems through an interface without needing more development. Also this type of server supports data in many languages.

3.4 Development and construction of the database

3.4.1 Metadata

Building the metadata is a necessity for the overseas Sinology database. Metadata are data about data. Their main functions are as follows:

- Discovery and identification: Identifying and discovering the digital information unit and collection. Dublin Core is a typical example.
- Cataloguing: Cataloguing the data unit in detail.
- Resource Administration: Supporting resource administration and access management including rights/privacy management, digital signature, seal of approval/rating, access management, payment, and accounting.
- Preservation and archiving: Describing the format of information, protection condition, and migration methods and support digital preservation. [5]

Metadata not only can describe the various kinds of information resources and the character and property of the data themselves but also can organize a mass of digital information into an organic structure so that the efficient and exact searches, materials-sharing, and data-management can be achieved. Therefore, metadata are necessary in developing the overseas Sinology database.

Metadata creation should conform to metadata standards as this is a precondition for search, exchange, and usage among different databases. A database should not become an information isolated island. The overseas Sinology database was designed to realize search and exchange among databases with related subjects. Metadata standards or specifications have been established for archives, geography, art, museum such as CDWA, GILS, FGDC, EAD, DC, and others [6].

However, developing metadata standards and specifications are different processes, especially for metadata which are suitable for specific subjects and formats. CALIS has set up a series of metadata specifications and a standard bibliographic description involving ancient books, stemma, rubbings, chorography, theses, and e-books [7].

Overseas Sinology is a database that includes information about people, organizations, and journals. There are no metadata standards in existence that describe the object data in the overseas Sinology database. Therefore, we must expand existing metadata standards to suit this database and create new metadata to describe and manage the objects. Thus, a new metadata system has been set up. The metadata about the overseas Sinology scholar database is based on the character metadata in the Chinese metadata standard framework, which has applied to Beijing University famous master collection. The works database references journal papers, e-books, and specifications and standards of the metadata in the digital library. The journal database metadata were newly created because there are no metadata standards to describe a whole journal, only a certain paper. The organization metadata was also newly created. In the process of creating a new metadata system, we considered not only the convenience of management, search, and exchange but also user demand. The bibliography items were established according to the metadata system. The following tables illustrate the information in detail.

Table 1. Descriptive metadata of a paper

core elements	Title, author, subject, secondary authors, description, date, type, format, identifier, language, related resources, right
other elements	degree

Table 2. Descriptive metadata of an e-book

elements	Title, author, subject, description, publisher, secondary author, date, type, format, identifier, source, language, related resource, time-space scope, right management

Table 3. Overseas Sinology metadata and cataloging items

Journal database	**Scholars database**	**Works & paper database**	**Institution database**
title of journal(Chinese)	name (original language)	type	name of institution(Chinese)
title of journal(original language)	name(English)	language	name of institution(original language)
title of journal(English)	name(Chinese)	discipline	name of institution(English)
country	male/female		country
language	homeplace		address
institution	nationality	author	telephone
newsroom address	birth date	editor	fax
telephone	year of die	translator	email
fax	degree	title of publication	website
email	discipline	specialty	
website	what kind of problem he/she studied	press	discipline
ISSN	title	press year	character of institution
discipline	duty	volume	tenet
date of start publication	institution belonged	author of paper	the date of foundation
evolution	address	translator of paper	evolution
period of press	telephone	title of paper	people
creator	fax	keyword	organization
	email	link to author database	main activity
	website	creator	publication
	resume		creator
	link to works & paper		
	creator		

3.4.2 Scope of the data

1. The major areas of the world: North America, Europe, the Middle East, East Asia, Oceania, 19 countries, especially adjacent countries.
2. A wide range of studies: not limited to scientific research organizations and scholars but also including academic publications and work in the research subjects.
3. Comprehensive information: for example, websites of research organizations, addresses of newsrooms, contact methods, resumes, and homepages for scholars. The information in the database is abundant and new.
4. Large time span: The earliest publication can be traced back to 1984, while the latest works are also embodied in the database. Information after 1980 accounts for a major part of the database[9].

3.4.3 Information resources

1. **Documents** - These documents in existence for a long period of time are the main reference resources for the database: *The Handbook of Sinology in America*, *The Sinology Scholar in Japan*, *Tibetan studies in China and Foreign countries*, *The Handbook of Sinology in Russia*, and *The World Overseas Sinology Scholars.* The Modern History Institute of CASS *China Study Overseas: Sinology Scholars in North America.*
2. **Internet information resources** - the websites of organizations, institutions, and individuals contain a variety of information. Internet information is continuously updated so that it is more comprehensive and newer than the information in print. However, its authenticity should be judged according to the creator. Therefore, the information collectors do painstaking work to analyze, select, and process data from the Internet, which may be absent, contradictory, and confusing. Email, fax, and telephoning are efficient means for authenticating information.
3. **Indirect methods** - information is obtained from universities, scientific research institutions, and guilds from local regions, nearby countries, and even the world. Information collectors attend seminars, dissertations, and lectures, talking with scholars, and obtaining first-hand information[10].

3.4.4 Data input

Manual data input of documents is better than scanning and using optical character recognition (OCR). The scope of the content is broader and more comprehensive than the existing materials. Information in the existing materials is old and insufficient for research. Scanning and OCR technology are not perfect. A great effort must be made to correct erroneous data. Therefore, we prefer manual input, which allows us to collect the newest information and compile it in a new system ensuring the veracity of the data.

3.4.5 Data verification

Verification of database quality must occur in every step of data processing. Because the users of database include everyone, especially scholars, and the data will be published in print, the quality of the data needs to be strictly controlled to assure its authentication and reliability.

During manual inputting, errors are inevitable. Incorrect keying and other kinds of mistakes must be corrected.

Methods of data verification include: repeated data verification, which deletes duplicate data in the database; visual method, which is an efficient way to find out a large number (75%- 85%) of mistakes; logic judgment, in which logic inconsistencies are verified; data typing, which will discover inconsistent data; and format verification.

Table 4 shows a verification entry [11].

Table 4. Data verification item

Category	Content
Title	The authentication of title, especially the translated title in Chinese If the spelling initially is wrong, assure the title is not repeated
Written/format	Spelling Punctuation Format
Acceptable table	The necessary entry
Missing material	The uncompleted content should complete or label the mark to faultiness
accordance	Check the accordance between Scholar database and Works & Paper database the accordance of data from different information source
value range and code	The value of data in the range; the character and code

Authentication and veracity of the data are very important because they affect the search results, research, and analysis. When this is completed, the data will finally be compiled and published. A good example of data verification is that translated names of the foreign scholars must be inspected by experts.

3.4.6 Data transmission

The final step is data transmission. The database administrator is in charge of uploading the data, and the database log will show the state of the transmission. Error data will be picked out automatically. Good transmission over the Internet is a good test for data transmission. After the transmission, the system will feed back information.

4 DATABASE APPLICATIONS

4.1 Background administration of the database system

1. Data management: including data upload, processing in batches, checking repeated data, backup, etc.
2. Customization of metadata: setup through the metadata template.
3. Data security: users can not casually alter, delete, or destroy the data. This can be insured by:
Individual view distribution – different users have different rights to operate the database, for example, the right to delete, read, and modify the data; and *user validation*: including Windows validation and SQL Server validation [12].
4. Encryption techniques: information can be read only after the rights are granted. The encryption techniques play an important role.

5. The release of the database and flexible setting.

6. Statistics: the database can provide log statistics functions. The statistical items include session time, frequency, IP, search, user information, etc.

Table 5. Log statistics

user	date	time	IP	How long	search	title	State of logoff
yq	2006-07-17	10:02	10.4.129.140	15’01”	5	2	
		Name of database			search	title	
		Overseas Sinology works & paper database			2	4	
		Overseas Sinology institution database			2	1	
		Overseas Sinology scholar database			1	1	

Other functions, not mentioned here, are available.

4.2 Database applications

4.2.1 An average application

1. **Log in**: IP login and user/password login
2. **Search**: basic search and advanced search.
   a. Basic search include key words, title, full-text, and subject.
   b. Advanced search includes every segment, Boolean search, and search on results.
   c. Search can be done in several languages.
3. **Browse**: the database has a map of the logical structure of the sub-databases. The database supports Unicode and multi-language display.
4. **Customization**: user can control the configuration, color, font, and icons.
5. **Http link**: by clicking on the link, the user can log into the website and email function;
6. **Search history**: the users can review what they have searched.

4.2.2 Special service applications

1. **Relation link**: after a search, related content is used for a new search point. The scope of information gets broader. Although it is a function of the relation link, a url need not be put into the address form, the relation link can be opened to log into the database and do the search directly. Thus the users’ information scope can be widened, and the limitations of the database alleviated.
2. **Encryption:** sensitive and critical data should be encrypted. Users who do not have the right to read the encrypted data just see – *This record is an encrypted one*.
3. **Sorting:** any data segment can be sorted in ascending or descending order
4. **Data mining**: there are four methods: statistical, robot-leaning, database, and nerve tracing net. After analyzing the character of the overseas Sinology database, the normal methods include a statistical method such as liner regression, multi-analysis, clustering analysis, differentiate analysis, relativity analysis and database methods such as multidimensional analysis, OLAP.

Analysis is usually done with SPSS software. Data can be downloaded from the overseas Sinology database and uploaded to the SPSS software. Using this software, we can mine a great deal of information and discover trends and relationships.

5 THE PRODUCTION OF THE OVERSEAS SINOLOGY DATABASE

At present, there are about 2,000 records of scholars in the Overseas Sinology Scholar Database. Aside from information about individual scholars, their work has also been collected and stored in the database. More than 500 records of organizations and institutes have been inputted into the Overseas Sinology Research Organization Database. More than 200 journal records about overseas Sinology have been deposited in the Overseas Sinology Journal Database (up to 2006-8-11). All the information is periodically updated and complemented.

The scope of the information exceeds the original areas and disciplines. Information comes from Canada, Australia, the UK, Germany, India, Australia, New Zealand, Singapore, Malaysia, and Korea along with the United States, Japan, and Russia, the pioneering countries in Sinology research. All these countries, not only adjacent countries in Asia, but also countries in Europe and on other continents, have an interest in Sinology research. The scope of the disciplines is expanding and the number of research organizations and institutions is increasing. Table 6 shows the distribution of organizations and institutions doing Sinology research. Figure 1 shows the distribution of the disciplines.

Table 6. Overseas Sinology institutions

Countries	Korea	United States	Japan	Germany	Russia	Australia	Canada	New Zealand
Number	106	92	56	54	39	38	26	9
Countries	Malaysia	Uzbekistan	Kazakhstan	Switzerland	Austria	India	UK	Singapore
Number	4	4	3	3	1	1	1	8

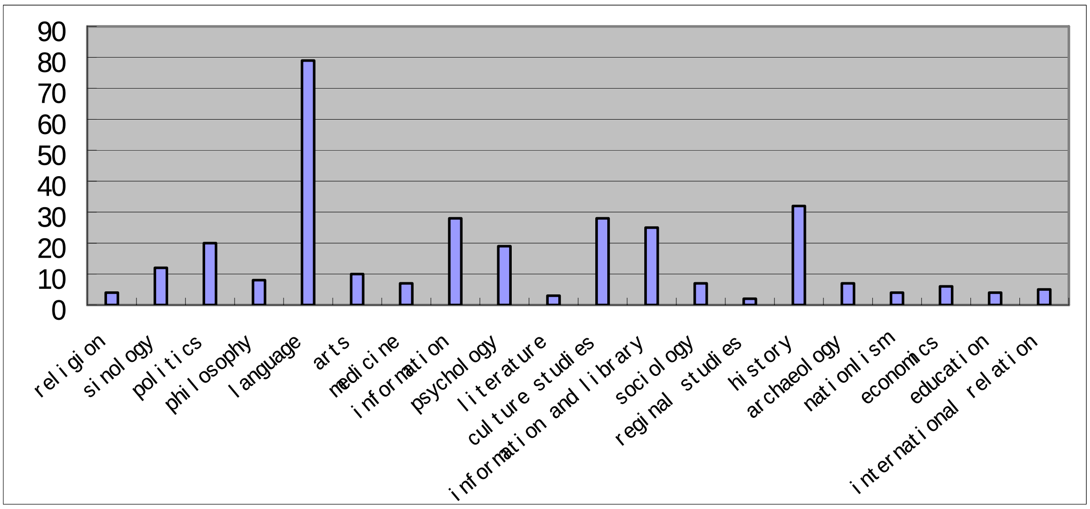

Figure 1. Overseas Sinology discipline distribution

From Figure 1, we can conclude that the hot points of Sinology include cultural research, politics, regional research, economic research, etc. Many institutions and organizations also undertake Chinese language learning activities.

In the Overseas Sinology Journal Database, journals containing overseas Sinology articles come from different countries, including Canada, the United States, Australia, Singapore, New Zealand, Holland, Germany, India, Korea, Japan, etc. The languages in the database are English, French, German, Korean, and Japanese. Some journals are published in both Chinese and English, some are published in several languages, and some are published in e-journals. All this information is in the database. Every record contains a title, ISSN, subject, publishing period, newsroom address, telephone number, fax, email, and website. This detailed information offers a convenience to users who can use this information to get more information, contact a newsroom directly, and easily access a website. The earliest start publication in the journal database is 1984, and the latest is 2002. Figure 2 shows the magnitude of change in the number of journals from the mid 19th century to 2002.

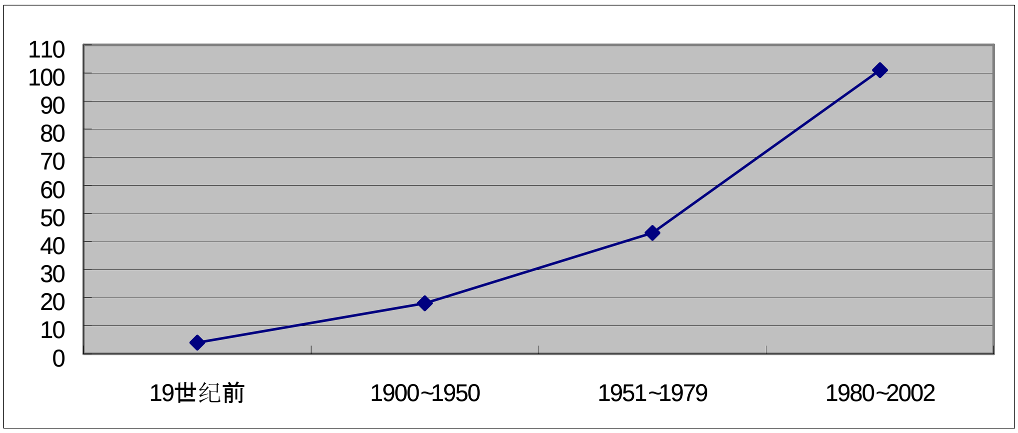

Figure 2. Magnitude of change in the number of overseas Sinology journals

Figure 2 illustrates that the development of journals about overseas Sinology was very slow before foundation of the new China. After the opening-up and reforms of 1980, Sinology research has developed rapidly with the result of an increased number of Sinology journal publications.

The search points of The Overseas Sinology Scholars Sub-Database are name, nationality, subject, discipline, and topic. Using the topic as a basic search point breaks out of the limited search such as on a name. Thus users can start with a problem about China or a subject to access related information, and different opinions on the same problem can be compared. This kind of search and analysis plays an important role evaluating and understanding new thought trends throughout the world.

The Overseas Sinology Database arouses great interest from Chinese and overseas scholars. It has great meaning for academic circles and government decision-making. Scholars from France, Korea, Japan, Viet Nam, Norway, and Russia have pared off with scholars from the Chinese Academy of Social Sciences and have give high praise to this database saying it is very practicable in the area of Sinology [14].

6 MAINTENANCE AND CONTINUOUS CONSTRUCTION

Maintenance of the database is a long-term and arduous mission. The work should be done, and the problems we should pay attention to are follows:

1. Update the data and compile new content. With the development of overseas Sinology, the database should reflect changes in development. New scholar information, new works, and new Sinology journals should be tracked and put into the database.
2. Enhance the application. Making the database known in governments and academic communities is further work that should be done. Applications can make the database play a realistic role and show its true value.
3. Integrate the information of this database into other information. Make the database easier and more convenient to use so that users can get more information from just one search or one accession.

There is much other work we could do. Building the database is just a start.

7 REFERENCES

[1] 20th Century International Sinology Research Library in Peking University. Retrieved from the WWW, December 19, 2007: http://www.cacl.org.cn/

[2] Oversea Sinology Research - a new developing way. Retrieved from the WWW, October 2, 2007: Http://www.zisi.net/htm/xzwj/zzhwj/2006-06-30-35157.htm

[3] Retrieved from the WWW, December 19, 2007: http://www.ccsa.cass.org.cn/

[4] Luo, Yan (2005) Construction of Distinctive and Special Subject Data Banks in CALIS and University Library. *Journal of Hunan Institute of Humanities Science and Technology (4).*

[5] Zhang, Xiaolin, The Standardization Framework of Metadata Development and Application. *Proceedings of the International Conference on National Libraries, Beijing.* The Library of Chinese Academy of Science.

[6] Report of a Comparative Study of Metadata Standards. *The Consortium for Research in Chinese Metadata Standards 2000 (12).*

[7] Xiao, Long & Chen, Ling The Standardization Framework of Chinese Metadata. *The Consortium for Research in Chinese Metadata Standards in the Digital Library of Peking University.*

[8] Retrieved from the WWW, October 2, 2007: http://162.105.138.23/bdms/index.asp

[9] Oversea Sinology Research since the Reform and Opening to the Outside World. Retrieved from the WWW, October 2, 2007: http://kyj.cass.cn/show_News.asp?id=2218

[10] Luo, Yan (2005) Construction of Distinctive and Special Subject Data Banks in CALIS and University Library. *Journal of Hunan Institute of Humanities Science and Technology (4).*

[11] Zhang, Kai (2004) Design and Realization of the Current Testing Implement for Input Data and the Reliability of Software. *Computer Development and Application (2).*

[12] Wang, Shan (1999) *Chen Hong, Principle of Database System*. Qsinghua University Press.

[13] Oversea Sinology Research since the Reform and Opening to the Outside World. Retrieved from the WWW, October 2, 2007: http://kyj.cass.cn/show_News.asp?id=2218

[14] A New Milestone for Oversea Sinology Research. Retrieved from the WWW, October 2, 2007: http://www.cass.net.cn/file/2005101847590.html