6-1-1993

The Development of Chinese Bibliographical Automation in China

Cheng Zhen

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

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/jeal/vol1993/iss99/4

This Article is brought to you for free and open access by the All Journals at BYU ScholarsArchive. It has been accepted for inclusion in Journal of East Asian Libraries by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
In 1991, in its Bulletin no. 5, the European Association of Sinological Librarians (EASL) published the article "Recommendations on automated cataloguing of Chinese books" to further the coordination in the field of nonromanized cataloging of Chinese collections in Europe. The main aim of EASL for cooperation was to keep open the free exchange of data, reduce possible technical obstacles to a minimum, and avoid some of the unnecessary duplication. To build a basis for cooperation, EASL decided to adopt the GB Chinese character code system as its most basic standard for the data exchange of Chinese characters. The reason for this adoption is that the main part of the Chinese collections in Europe have been and will be obtained from mainland China, the "country of origin", and so will the catalog data.

Though the development of automated bibliography in China itself is not yet as advanced as could be hoped, it has indeed made great progress in the past decade.

Any successful computerized Chinese bibliographical system should have three elements: (a) a system of processing Chinese characters, (b) Chinese MARC data, and (c) an application software system. China has made great strides forward in all three fields, while European Chinese librarians are making every endeavor to find a way to create the automated bibliography of their Chinese collections.

This paper aims to offer some information which, I hope, will be of help to my European colleagues as they decide which way to proceed towards the automation of Chinese collections.

1. The Processing System of Chinese Characters

The computer processing of Chinese characters has long been the fundamental hurdle to making an automated bibliography of Chinese collections. To make computer processing of Chinese characters possible, a great deal of effort has been expended to accomplish this tough task. Nowadays, several different code systems are being used in various databases in different countries and areas of the world. But these databases are cut off from each other by using these different code systems.

Considering the need for exchanging records among European Chinese libraries and for deriving an automated catalog from mainland China, EASL has decided to adopt the GB system as its basic standard. Since the publishing of the GB code of Chinese characters primary set in 1981, China has made great progress.

1.1 BG Code of Chinese Character Set
On May 1, 1981, the "Code of Chinese Graphic Character Set for Information Interchange (Primary Set) GB 2312-80" was published in China. Only 6,763 Chinese characters were included in this primary set. This was due to the consideration of helping reduce the cost of setting up a Chinese character processing system so as to popularize the use of the computer in China. It was wise that the number of Chinese characters in the Primary Set had been controlled at this level while there was only a very limited amount of memory available in a microcomputer. According to the statistics of a Chinese character usage frequency survey conducted in 1975, the coverage rate in terms of the use of the 6,700 or more characters of the Primary Set was over 99.99%. Obviously, the 6,763 Chinese characters collected in the Primary Set was insufficient for cataloging Chinese collections in Chinese libraries.

To meet both the needs of those users who use more characters than those contained in the primary set and the needs of the geographical areas where the traditional form of Chinese characters are used, five supplementary sets were published since about 1984. The amount of characters collected together in each set are shown below:

<table>
<thead>
<tr>
<th>Set Name</th>
<th>GB Code</th>
<th>Number</th>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Set</td>
<td>GB 2312-80</td>
<td>6,763</td>
<td>Simplified</td>
</tr>
<tr>
<td>Supplementary Set 1</td>
<td>GB 12345-90</td>
<td>6,866</td>
<td>Traditional</td>
</tr>
<tr>
<td>Supplementary Set 2</td>
<td>GB 7589-87</td>
<td>7,237</td>
<td>Simplified</td>
</tr>
<tr>
<td>Supplementary Set 3</td>
<td>GB 13000 (?)</td>
<td>7,237</td>
<td>Traditional</td>
</tr>
<tr>
<td>Supplementary Set 4</td>
<td>GB 7590-87</td>
<td>7,039</td>
<td>Simplified</td>
</tr>
<tr>
<td>Supplementary Set 5</td>
<td>GB 13000 (?)</td>
<td>7,039</td>
<td>Traditional</td>
</tr>
</tbody>
</table>

After publishing these five sets, the GB set family has a collection of 42,181 Chinese characters in total. Some of the characters have both the simplified form and the traditional, complex form. After reducing these two forms to count as one character, the GB set family has about 28,000 independent Chinese characters, which is quite enough for library cataloging of Chinese books. Most of these characters are rarely-used or occasionally-used characters.

1.2 Electronic Dictionary of Chinese Character Attributes

The Electronic Dictionary of Chinese Character Attributes is another very important facility for automated cataloging of Chinese collections and Chinese information processing. In 1985 the Database of Chinese Character Attributes has been developed by the National Library of China (NLC), which comprised twenty-five attributes of 6,763 Chinese characters in the Primary Set. Based on this database, the Electronic Dictionary of Chinese Character Attributes was built in 1990. The Dictionary has more than 40,000 Chinese characters taken from the six character sets. Supported by a software system, this electronic dictionary possesses the following basic functions:

A. **Chinese Character Sorting**  Considering that several authoritative rules for sorting Chinese characters are widely used in word processing, this dictionary enables a computer-
based Chinese information processing system to arrange Chinese characters in (a) component and radical order, (b) Chinese pinyin order, (c) stroke number and stroke form order, and (d) the four-corner system.

B. Phonetic Conversion This system implements automatic conversion between a Chinese character and its phonetic transcription so that the system can meet the needs of the romanized Chinese catalog database of overseas users. That will save the cost in indexing, inputting, checking, and updating pinyin data, which is always subject of error. With the help of this function, the romanized catalog data have been created automatically in the Chinese MARC published by NLC.

C. Variant Form Connection As a matter of fact, the simplified form, the traditional complex form, and variant forms of Chinese characters are now all in use in different areas: mainland China, Taiwan, Hong Kong, and other countries. In order to enhance the completeness of the retrieval system, to obtain information whether using the simplified form, the traditional form, or the variant form, this function offers the linking up of the simplified form to its traditional and variant forms.

D. Code Conversion For the sake of exchanging Chinese information among computer systems which adopt different Chinese character code sets, this function builds a relationship between the GB code system and the CCCII code system, thus making it possible to use CCCII-based MARC data in a GB-based computer system.

E. Other Applications Because this dictionary contains all the various attributes of Chinese characters' form, phonetic, and code, it facilitates research in the characteristics and the patterns of Chinese characters.

This system is applicable in macrocomputers, medium computers, minicomputers, and microcomputers. Since 1986, this attribute dictionary and its software have been used widely by hundreds of domestic users. In 1990, the research project on the Chinese Character Attribute System was awarded the National Science and Technology Progress Prize.

1.3 Multilanguage Processing

Since most libraries have their collections in more than one language, we have to keep an eye on the automated processing of multilanguage and multiscript materials as we try to find the solution to the processing of Chinese information.

A software package has been developed by NLC in its NEC ACOS mainframe with the capability of processing Chinese, Japanese, "western language" (the Latin family) and Russian. The multiplanar-invoking technique is used to call up a different character set when necessary, but this is by no means the final solution to multilanguage processing.
To meet the challenge of the multilanguage environment, many multilanguage processing systems (such as CJK) have been used in different areas and different countries. Though these systems have the ability in some degree to process more than one language, the use of different character code systems makes exchanging data difficult, if not impossible, between different systems.

In view of the fact that what we face today is not only the challenge of multilanguage, but also the challenge of multicode systems, the International Standards Organization has issued a standard document — ISO 10646 — to set up a universal language character set. In the CJK section of this character set, 20,902 Chinese characters are involved, of which about six hundred characters are used only in Japanese and fifty only in Korean, the others being used in Chinese characters. There are also one thousand characters in English, Latin, Cyrillic, Greek, Arabic, and so on in this universal character set.

Based on this ISO standard, some United States computer companies have started developing a new generation of computer system especially for the use of cataloging multilanguage materials in libraries. It is expected that the new computer system will be commercially available in one or two years. That will lead to the final solution of multilanguage processing and also the final solution of Chinese information exchange in different databases.

2. Chinese MARC

2.1 Chinese MARC Format

The National Library of China launched into its plan of library automation, which is bibliographic-based and Chinese-oriented, as early as 1975 while the whole country was still in the chaos of the "Cultural Revolution". In 1980, NLC introduced Library of Congress (LC) MARC and built a simulation system with the functions of database creation, data retrieval, and so on. Based on a series of research works, NLC began its project of developing Chinese MARC (CNMARC). In December 1986, after research into the experiences of automated cataloging in other libraries throughout the world, the Automation Department of NLC drafted the Chinese MARC Format. In 1989 this draft format was revised according to the newly published UNIMARC and the first edition of the Chinese MARC Format was officially published in 1990.

Based on a series of international standards as well as national standards, this document stipulates the tag, indicator, and subfield identifier of a catalog record, and also stipulates the logical and physical formats with which the catalog records are recorded on tape, disk, or other media, and the format of its content designator. The main features of this MARC format are the following:

(a) Complying with the designing principles of UNIMARC and of the international standards concerned, it is easily applied in international exchanges of catalog data.
Considering the distinguishing features of Chinese cataloging data and the processing of Chinese characters, this format adds certain fields, subfields, and content designators to meet the needs in cataloging Chinese collections.

With reference to the international standards, this format stipulates the identifiers for different character sets and the rules for switching between character sets. Thus it meets the needs of cataloging no matter what characters from different sets are needed for the cataloging data, whether they be one-byte characters or two- or three-byte characters that appear alternatively.

Since 1988 NLC has cataloged Chinese books published in mainland China conforming to the Chinese MARC format and put its MARC data on sale domestically in 1990. The export of CNMARC data to overseas countries was approved by the Ministry of Culture one year later.

2.2 Chinese MARC Data Center and Its MARC Products

Aimed at making CNMARC products commercially available to libraries at home and abroad, the Chinese MARC Data Center was set up in January 1991 under the New Technology Development Company of the NLC. Now there are four departments under this center including the Data Collection Department, the Cataloging Department, the Technical Development Department and the Issuing Department; altogether the Center has about sixty professional staff members.

The main products of the Chinese MARC Data Center are as follows:

- **Chinese MARC** Issues every two weeks with 1,200 records in each issue. Now up to 100,000 MARC records of the books published since 1988 are available in NLC's Chinese bibliographical database.

- **Catalog of Current Chinese Serials** This catalog is a part of CNMARC, updating over 10,000 titles of newspapers and periodicals published from 1988 to 1991 in mainland China, Taiwan, Hong Kong, Macau, and several Chinese newspapers published overseas.

- **Chinese Character Attribute System** See paragraph 1.2 for details about the system.

- **Library Microcomputer Management System** This is a microcomputer-based software package, catering to the needs of medium and small libraries. See paragraph 3 below for details.

- **Chinese Bibliography Authorities Database** Established on the basis of CNMARC format, the Authorities Database contains personal authors, family...
authors, group authors, conferences, geographical names, uniform titles, series titles, and subjects. As a part of this database, the Authorities File of Ancient Chinese Authors is now being compiled and contains over 40,000 personal and group authors from ancient times to the period of the Republic of China (1911).

3. Application Software System

Generally speaking, a Chinese computer software system can be divided into three layers as shown in the figure below. At the top is the operating system (such as CCDOS); under this layer is the Chinese character font bank system (including the code and dot-matrix font bank); corresponding to the font bank system, the Chinese Character Attribute System is installed in the same layer to support the computer processing of Chinese characters. Under this layer is the application layer, where various kinds of computer programming languages, application programs or application software packages are installed.

<table>
<thead>
<tr>
<th>Operating System</th>
<th>Font Bank System</th>
<th>Attribute System</th>
</tr>
</thead>
<tbody>
<tr>
<td>(CCDOS, CUNIX, etc.)</td>
<td>(code, font bank)</td>
<td>(Attribute database, software)</td>
</tr>
<tr>
<td>Application Layer</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NLC is now developing a comprehensive management software package in its NEC ACOS-603 maxicomputer system. The main part of the software package, including acquisition, cataloging, retrieval, authority control, etc., is already in operation.

Concurrently, the Chinese MARC Data Center has developed a microcomputer library management system, catering to medium and small libraries. This software package contains acquisition, cataloging, retrieval, product output, database maintenance, circulation, and serials management subsystems and can be run on IBM microcomputers (286 or over) or their compatibles.

This system allows users to select catalog data from CNMARC and send them to the user's database directly. The retrieval subsystem provides seven access points, among which four access points allow logical retrieval. The output subsystem is applied in printing new books bulletins, catalog cards, bibliographical monograph titles, or the catalogs of large collections.

This microcomputer-based system is designed to accord with the situation of most local public libraries in China. An advantage with this system is that no programming work is needed, thereby saving the users' work and time in programming tasks which are difficult for most librarians. One disadvantage is the system's lack of flexibility.
4. Suggestion

The endeavors that China has made in the Chinese bibliographic automation field, as mentioned above, are a part of a worldwide effort. As a "country of origin," China plays an important role in this effort. China's achievements in the processing of Chinese characters, research on Chinese character attributes, the production of Chinese MARC data, and so on, have provided an essential prerequisite for Chinese bibliographical automation. Chinese libraries both at home and overseas will benefit from these fruits. Unfortunately, China has paid such little attention to communication with overseas Chinese libraries in the field of automation that very little is known by overseas colleagues about what China has achieved in these past few years. The China MARC Data Center of NLC is now looking for overseas partners to promote the spread of Chinese MARC throughout the world. Obviously, for European Chinese libraries, obtaining MARC data from China will save them work in cataloging. Communication and cooperation with China is therefore important for the exchange of information and ideas and to derive the proper MARC product from China.

My suggestion at this stage is that EASL establish a closer relationship with the Chinese MARC Data Center to share common interests.

Before proper cooperative action can be taken, we will need a reliable channel of regular communication, so that EASL can obtain news about ongoing progress in China and can obtain the required documents and materials issued by NLC about CNMARC. On the other hand, the Chinese MARC Data Center can learn more about the needs of European Chinese libraries and about the environment for the automating of Chinese catalogs in Europe. Through this channel of communication, China can exchange information and ideas not only with individual libraries but with European Chinese libraries as a whole.

Though NLC has established cooperative relations with some libraries, the further cooperation between the Chinese libraries in Europe and the Chinese MARC Center should be conceived on a European scale. As Mr. Howard Nelson, librarian of the British Library Reference Division, said ten years ago in his paper entitled "Chinese Catalogue Automation in the UK: And in Europe?": "If progress is to be made at all, it should probably now be conceived on a European scale, no longer on the basis of a single country – let alone a single library."

REFERENCES


