Chinese MARC (Taiwan) and Its Bibliographic Database

by

Ching-Chen Anthony Mao
Fu Jen Catholic University
and
Ching-fen Frances Hsu
National Central Library
Taipei, Taiwan

Abstract

The Chinese MARC Format (CMARC) was first published in 1982, and is still the most widely used machine-readable format among libraries in Taiwan. Besides the background and current application of the CMARC, this paper describes two subjects: how CMARC adopts and differentiates from UNIMARC , and Chinese character internal encoding systems. Steps are taken to bridge between CMARC and other data formats, for instance, recent development of CMARC3-XML Schema and MARC 21 to CMARC3 mapping table. The next major task will be how to prepare for the challenge of international bibliographic data exchange.

The origin of CMARC

The introduction of the MARC format by the Library of Congress in the 1960s pushed the library towards an era of computerization. Computerized bibliographic records not only enhanced the data retrieval and storage but also served as the basis of data exchange. In the mid-1960s machine-readable formats were developed almost concurrently but separately by the Library of Congress and by the Council of the British National Bibliography. Though there was cooperation in developing MARC II in 1968, in order to meet the needs for different cataloguing practices and requirements, various MARC format, such as USMARC, UKMARC and INTERMARC, emerged in the 1970s.[1] The creation of UNIMARC as an international machine-readable format provided libraries with the solution to the problem in data exchange between different MARC formats. Influenced by the development of the MARC format and considering both local application and international data exchange, the Chinese MARC Format was thus created.

The Library Association of China and the National Central Library (NCL) jointly established the Library Automation Planning Committee (LAPC) in 1980 to improve library and information management and services. One of the objectives of the automation planning by the LAPC was to develop the machine-readable format as the standard for cataloguing Chinese publications. The Chinese MARC Working Group (CMWG) was formed under the LAPC to design a machine-readable format that would not only process Chinese materials but would also conform to the standards for international data exchange. The decision was made to take UNIMARC as the model for designing the CMARC and USMARC as a major reference.[2]

In 1981, Chinese MARC for Books was published mainly for processing monographic materials. The CMWG continued revising CMARC in reference to the new version of USMARC Formats for Bibliographic Data and UNIMARC so as to enhance the feasibilities of the CMARC format. In 1982, the First Edition of Chinese MARC Format (CMARC) was published with elements to process serials, maps, music and audio-visual materials in addition to monographs. With the help of the Institute of History and Philology of the Academia Sinica and the Division of Special Collections of the NCL, CMARC fields dedicated to Chinese rare books and rubbings were added to the Second Edition of CMARC published in 1984.

When the Third Edition of CMARC was published in 1989, libraries in Taiwan were in the phase of automation. Being promoted as the standardization of library automation, the Third Edition of CMARC was the first MARC format used by many libraries for its capacity to cover almost all the materials held by a library at that time. Even after the Fourth Edition was published in 1997 and the Version 2001 in 2002, the Third Edition of CMARC is still the most widely used MARC format among libraries in Taiwan.[3]

State of the Art

Features of CMARC

The basic structure of CMARC involves record structure, content designation and data content just like UNIMARC or other MARC formats. The features of CMARC are:

- developed according to the characteristics of Chinese materials
- applicable to materials in various languages
- applicable to various types of materials
- implemented with ISO 2709/CNS-13148 for bibliographic information interchange on magnetic tape
- implemented with Chinese Cataloguing Rules [Taiwan] for Chinese materials and AACR2 for western materials
- implemented with Chinese Character Code for Information Interchange (CCCII) or Big5 for Chinese data
- in conformity with international standards for geographic areas codes, time period codes, country codes, language codes, cartographic materials codes.

A Concise CMARC is designed for small libraries[4].

In order to comprehensively describe all types of Chinese materials, data fields and codes dedicated to specific type of Chinese materials, which are not defined in UNIMARC, are added to CMARC.

For example:

- Field 100 General Processing Data $a/26-29 Character Sets — use code "90" for CCCII, "91" for Big5, "92" for CNS-11643, "93" for GB

- Field 128 Coded Data Field: Music Performances and Scores — at subfield $a, use additional codes from "ya" to "yz", etc. for Chinese music performances; and at subfield $c, use additional codes from "te" to "tk" for Chinese strings, from "wj" to "wr" for Chinese woodwinds
- Field 129 Coded Data Field: Rubbings $a/0-6 — include codes for Type of Rubbings, Method of Production, Forms of Materials(two bytes), Style of Calligraphy, Style of Character and Colour of Ink and Water.

Development of CMARC

The CMARC format is developed based on the model of UNIMARC. It is therefore important to keep the harmonization with international standardization. However, since the CMARC format is meant for use in libraries in Taiwan, factors such as librarians' adaptability and implementation in library systems need to be taken into consideration.

During the process of modifying CMARC in the late 1990s, opinions from experienced librarians and library system vendors as well as library scholars were invited. The discussions on modification resulted in replacing the Linking Entry Block (4XX) with equivalent Related Title Block (5XX). Both librarians and library vendors will need crucial adjustments if they want to fully comply with the modifications. In the current situation libraries adopt the new version of CMARC in different ways. Some select new fields for certain purposes such as for cataloguing new types of medium.

The task of modifying CMARC will continue to be made under the principles of maintaining structural integrity and embedding elements from current development of both UNIMARC and MARC21. It is foreseeable that, in the process of future modification, there will still be debates over issues regarding the MARC structure and practice in the library. Hopefully the next version of CMARC will emphasize setting long-term strategies to extend its feasibilities for practical requirements in library and for maintaining the stability of CMARC structure.

CMARC3 XML schema/DTD

The MARC format which conforms to ISO 2709/CNS-13148 is considered to be the standard among most libraries, whereas the development and utilization of XML has become a trend for data processing and transportation outside librarianship. In order to increase the possibilities of data sharing, Dr. Shien-Chiang Yu from the Shih-shin University launched in 2004 a research project funded by the NCL to construct CMARC XML[5].

Because it contains document type definition and follows the standard format in data input, XML becomes an ideal tool for data exchange or transformation across systems. Compared with XML, the MARC format which conforms to ISO 2709/CNS-13148 cannot recognize the MARC type; nor can the content be directly presented on the web. The drawbacks of the MARC format limit its application to automation systems.

The project includes analysis of both foreign and domestic methods of schema formation which adopt XML as the data format for bibliographic data exchange with references to interrelated definitions and contents. As part of the project, a program is developed to convert ISO2709/CNS- 13148 files to and from XML documents based on XML Schema. The documents of CMARC3 XML schema/DTD and the conversion software could be downloaded for trial registration.

CMARC3 to MARC21

Among the libraries with comparatively large holdings, CMARC is still the most widely used MARC format for cataloguing Chinese materials in Taiwan. On the other hand, during the past two decades the libraries in Taiwan used extensively bibliographic resources in USMARC for materials in Western languages. Since the majority of Western language collections in the libraries are in English, cataloguers depend a lot on deriving bibliographic resources in USMARC/MARC21 provided by OCLC and ITS MARC.

In order to avoid data loss during MARC conversion, many libraries use the CMARC format for Chinese materials and USMARC/MARC21 for materials in Western languages. For those libraries that need to derive resources in USMARC/MARC21 but use only CMARC or vice versa, in-house programs will have to be developed to convert data into the needed MARC format. Most of the MARC conversion programs are developed and built within the library system. It is important to ensure that the conversion programs are designed based on the same standard.

In 1992 the Ministry of Education funded a project to develop specifications for the conversion of bibliographic records in CMARC to and from USMARC. The members of the project were experts in the MARC format and librarians experienced in using CMARC or USMARC. The project resulted in MARC field mapping in tabular form in a two-volume set published in 1993, one for converting bibliographic records in CMARC to USMARC and another from USMARC to CMARC. Besides, the project also includes a suggested prototype for designing conversion programs and related technical documents.

To reflect the current usage of MARC21, the NCL just completed the conversion specifications from CMARC to MARC21 in April this year. The specifications are established in reference to UNIMARC to MARC21 conversion specifications (Version 3.0) and reviewed by library scholars. The specifications are expected to enhance the resource sharing of bibliographic records in Taiwan and also to improve international bibliographic exchange such as uploading data to OCLC.

NBINet union catalogue

One of the goals of developing the CMARC format is to foster an online union catalogue. The NCL launched the National Bibliographic Information Network (NBINet) in 1991. The current system started its operation in 1998 to cope with bibliographic records in various MARC formats and Chinese internal codes contributed by member libraries. In the 1990s, besides CMARC, USMARC became popular especially for cataloguing materials in Western languages. As for Chinese internal codes, Chinese Character Code for Information Interchange (CCCII) and Big5 are the most widely used Chinese internal codes among libraries in Taiwan. Due to the divergent developments of library systems used by cooperative libraries, the MARC format and Chinese internal code are always the major concerns for establishing a union catalogue in Taiwan.

The NBINet system is able to store bibliographic records in multiple MARC formats which conform to the ISO 2709/CNS-13148 standard but the Chinese internal code used for input currently has to be CCCII. The internal code will be converted to Unicode in the near future. The bibliographic files provided by the member libraries could be in any MARC format with CCCII, Big5 or Unicode. All these files will be converted into CCCII before

loading into the database. To satisfy needs for different data formats, the system is able to output bibliographic records in certain MARC formats and internal codes selected by the member library.

Multiple MARC formats

NBINet currently has 77 member libraries. Among the member libraries, 67 of them use CMARC to catalogue materials in Chinese, Japanese and Korean; 32 out of the 67 libraries use only CMARC. Ten out of 77 libraries use only USMARC/MARC21. 35 out of 77 libraries use CMARC for CJK materials and USMARC/MARC21 for materials in other languages. It is likely that the majority of the collections in almost all libraries are in Chinese. Since the NCL has most of the Chinese materials published in Taiwan, most libraries will follow the MARC format used by the NCL for cataloguing Chinese materials. On the other hand, the bibliographic resources for materials in Western languages, especially those in English, are almost all in USMARC/MARC21. Libraries tend to use USMARC/MARC21 as well as CMARC to avoid the data loss of MARC conversion.

The advantages of using multiple MARC formats in a union catalogue are: (1) the coverage of bibliographic resource is extended without being limited to a single MARC format; (2) no effort is spent on MARC conversion to pre-process the input files; (3) there is no data loss if a record is input and exported in the same MARC format. Nevertheless, there are still disadvantages: (1) there are duplicate records for the same work but in different MARC formats; (2) libraries have to check the MARC format before deriving records; (3) data loss caused by MARC conversion is inevitable if a bibliographic record is exported in a different MARC format from its original one.

Issues of Multiple internal codes

The diversity of Chinese internal codes has long been a problem for library systems used in Taiwan. The commonly used internal code sets among libraries are CCCII (around 54,000 codes) and Big5 (around 13,000 codes). The type of internal code implemented in the library system will affect the quality of the processed bibliographic records and patron records. Among the 77 NBINet member libraries, 38 of them use CCCII, 32 libraries use Big5 and currently only 7 libraries use Unicode. Libraries using CCCII have more choices of characters than those who use Big5. However, CCCII is applied only to a particular library software in maintaining bibliographic or patron records. A lot of codes are still unable to be displayed with Web OPAC in either Big5 or Unicode. On the other hand, the Big5 system can display exactly what is input in the bibliographic record but librarians will usually encounter the problem of insufficient characters[6].

Actually for either type of code designation, new codes have always been demanded by librarians. Unfortunately, there is no organization responsible for regular maintenance and the libraries just can not wait for the long process of assigning new codes. In order to solve the problem of insufficient characters, different vendors utilize user-defined areas in different ways which results in difficulties for data exchange. With more than 70,000 CJK codes, Unicode is no doubt a solution to the chaotic situation.

Whether to convert to Unicode or not, depends partly on the system vendors and partly on the standardization for conversion. If a vendor decides not to spend unaffordable efforts to do code conversion on the current system, the library will need to evaluate whether to keep the system or to take alternatives to use Unicode. Normally the alternatives will always bring up the budget issues, which need long-term planning. The standardization for conversion would help to avoid data loss and incorrect conversion. Converting data from Big5 to Unicode is expected to cause no problem since Unicode is likely to include all the characters in Big5. To convert from CCCII to Unicode requires careful preparation because the CCCII code set has the feature that multiple codes are mapped to an identical character for structural arrangement.

The unofficial Unicode Workgroup formed in 2004 for library purposes and hosted by the NCL has the following aims[7]:

- to establish a code mapping table as a standard for data conversion from CCCII to Unicode;
- to establish a code mapping table for data conversion from Unicode to CCCII;
- to establish a preferred CCCII listing for characters with multiple mapped codes;
- to maintain the modification and extension of the above mapping tables.

The Workgroup finalized two-way mapping tables including more than 50,000 mapping sets from CCCII to Unicode and more than 46,000 sets from Unicode to CCCII. The current mapping tables should cover almost all characters that are frequently used. Additional mapping sets for rarely used characters will be added in the next version. These tables are not only used to prepare for the Unicode environment but also to provide data exchange standards for the interim while CCCII is still used among libraries.

Conclusion

One of the missions of a library is to preserve the cultural legacy reflected in various forms of publications. Different MARC formats and language encoding systems are developed as standardized tools to properly record and store the publications held by libraries in different countries. Mutual respect is needed for diverse standards representing different cultures. Although UNIMARC as well as Unicode is aimed at bridging different standards, they are still unable to fully encompass all elements in CMARC or all Chinese characters. The best solution to manage Chinese materials is to improve the current standards and to maintain compatibility with other languages.

References

1. The UKMARC Manual: Preface, http://www.bl.uk/services/bibliographic/marc/marcintro.html, accessed 27th June 2006
2. Chinese MARC Working Group, Library Automation Planning Committee, "Preface," Chinese MARC Format for Books (Taipei, Taiwan: Library Association of China & National Central Library, 1981), pp. iii-iv.
3. Mao, Ching-Chen, The Compatibility of CMARC [in Chinese], Journal of Educational Media and Library Sciences, 35(4): 310 - 337, 1998. Huang, Mei-Lien and Huang, Wen-Yu, A Comparative Study of the CMARC3 and CMARC4[in Chinese], Bulletin of Library and Information Science, 39 (Nov. 2001): 94- 108.
4. Chiang, Hsiu-ying, "Introduction to MARC format" [in Chinese], Library Association of China Workshop on Management of Library Resources, 26-31 July, 1999 (Taipei: National Central Library, 1999), pp. 20-21.
5. Yu, Shien-Chiang, MARC XML Schema/DTD report [in Chinese], Taipei, National Central Library, 2004.
6. Chinese Code Introduction, at CNS 11643 Full Character Repository, http://61.60.106.73/eng/word.jsp, accessed 27th June 2006.
7. Unicode Workgroup [in Chinese], http://unicode.ncl.edu.tw/, accessed 27th June 2006.