CATALOGING CJK ONLINE: THE ASIA EXPERIENCE

Jay Lee*

ABSTRACT

This article is to study the development of college and university libraries during the past 35 years in Taiwan. For looking back the development before 1982, the author cited some statistic results from the investigation of university and college libraries in 1972 and 1982.

In the meantime, this article attempts to analyze the current development of university and college libraries, and compares the results with those obtained in the investigation of 1972 & 1982. Purposely, it tries to find out some problems existing in our college and university libraries.

In conclusion, the importance of university and college libraries has been recognized, but there are some efforts we should pay to improve them, e.g., the professionalization of personnel, the promotion of service quality, and the execution of functions for instruction and supporting research.

I. INTRODUCTION

ASIAN SHARED INFORMATION AND ACCESS (ASIA) was funded in 1981 with a Library Services and Construction Act (LSCA) grant. The purpose was to establish a cooperative center for selecting, purchase, and cataloging Asian language materials in Chinese, Japanese, Korean, (CJK) and Vietnamese. To meet the needs of the participating libraries, ASIA has expanded its services to include technical processing and retrospective conversion.

*Assistant Director, Asian Shared Information & Access.
ASIA's current retrospective conversion projects for San Francisco State University and the California Statewide Data Base, believed to be the first two Asian language conversion projects in North America, are scheduled for completion in June and September 1989 respectively. So far 6,000 records from these two institutions have been converted and entered into the OCLC Online Union Catalog.

ASIA is still supported in part by the same LSCA grant. Each year ASIA serves 30 to 40 library institutions. So far a total of 60 library institutions have participated in this cooperative program — 50 in California and 10 out of state. Among ASIA's participants are 5 academic, 2 law, 2 special libraries, and 4 agencies — California Statewide Database, California Opportunities for Reference Excellence (CORE,) Cooperative Library Agency for Systems and Services (CLASS,) Library Association of Portland, and North Carolina Foreign Language Center.

Asia currently has a total staff of 18. Among the 9 professional staff, 7 hold the MLS degrees. All ASIA staff are native speakers of the respective languages. This staff represent one of the most established human resources for Asian language contract projects.

II. CATALOGING ACTIVITIES

In September 1983, ASIA started to catalog online on the RLIN CJK system. In September 1987, ASIA installed the OCLC CJK system. ASIA currently operates 4 OCLC CJK, the largest chain of the OCLC CJK, and 2 regular OCLC terminals. RLIN's new Multi-Scripts Workstation (MSW) has offered ASIA a cost-effective system to resume its services to the RLIN users. Whenever requested to provide services on the RLIN system, ASIA is ready to install the RLIN MSW.

As of March 1989, ASIA has purchased and cataloged over 22,000 titles and 170,000 volumes.
In 1986, ASIA's 8,000 original records cataloged on the RLIN CJK were tape-loaded into the OCLC database. As of January 1989, ASIA contributed a total of 10,988 original CJK records into the OCLC database.

ASIA's current retrospective conversion projects for the California Statewide Data Base and San Francisco State University are scheduled for completion in June and September respectively. So far 6,000 records from these 2 institutions have converted on the OCLC CJK and entered into the OCLC database.

III. OBSERVATIONS ON PRODUCTIVITY

During the 4 years of cataloging on the RLIN CJK, ASIA used one terminal and cataloged approximately 2,000 original titles per year. In the last 12 months, ASIA cataloged approximately 3,000 original titles on the OCLC CJK. The latter output amounts to a 50% increase from the former. During this 12 month period, ASIA had an increase of 2.5 times CJK terminal time. And during the same period, ASIA staff basically operated on the same schedule and workflow. This disproportionate increase of productivity, therefore, is likely correlated to the efficiency of the two CJK systems.

While, the OCLC CJK is clearly more user-friendly than the first generation of RLIN CJK, the latter is likely more efficient than the former once the learning curve is achieved. Many reasons may affect the operating efficiency. As follows are the most likely ones:

1. The first generation RLIN CJK runs on the AT configuration which is several times faster than that of the PC upon which the OCLC CJK runs.
2. The OCLC CJK users share the online access system with more than 10,000 OCLC terminals. The telecommunication "traffic jam" forces the users to wait for the stop-and-go "traffic lights."
3. The OCLC online database is many times larger than that of RLIN. A larger record universe takes longer time to search. Obversely, a larger database is an advantage because it contains more matching records and reduces the rate of the time-consuming original cataloging. The OCLC database, however, has less CJK records than does the RLIN database. The net result is that OCLC CJK users have to search a much larger universe to find less matching records.

4. The RLIN online system provides a more advanced search key structure. While the current OCLC search keys are very easy to learn and adequately efficient, especially in view of its database size, searching on the OCLC online system and evaluating the matching records are more time-consuming than performing the same on the RLIN CJK.

The above observation does not take the cost factors into consideration. As RLIN has introduced the new CJK which costs less than 25% of the original version and has many more user-friendly features, I expect it to be more cost-effective. Meanwhile, to my knowledge, OCLC is testing several overdue enhancements for its CJK. Some of these enhancements are discussed in the next section. Not until OCLC introduces a new version of CJK (more than just enhancements) the RLIN CJK will likely be more efficient. As a user of OCLC and a potential user of RLIN, ASIA looks forward to the continued enhancements and improvements from both online systems.

IV. CATALOGING AND TECHNICAL PROCESSING ISSUES

ASIA staff share the following constraints with other OCLC CJK online cataloging users:

1. CJK records have a low hit rate on the OCLC Database.

The hit rate is lower than 5% for retrospective titles (published in 1978 or older) and the most current titles. The overall hit
rate is approximately 20%. As the OCLC CJK system has been in use for only two years, and most of the major East Asian libraries use the RLIN CJK, the low hit rate it expected to continue until the CJK records exchange program between the two online systems is implemented. Since this agreement only calls the exchange of the CJK records created from December 1988 on, the low hit rate is expected to continue for some time to come. ASIA looks forward to a complete CJK record exchange between the two organizations.

2. Searching and evaluating the matching records on the OCLC online system are time-consuming.

There are many ambiguous situations that warrant a consensus of the East Asian libraries. For instance, if all the bibliographic elements of 2 records match except that the printing dates vary for, say, more than 5 or 10 years, or if one record contains the edition statement while the other does not, should these records be considered matching? Since both OCLC and RLIN periodically tape-load the LC records, each organization should have developed the matching criteria to weed out the duplicate records. ASIA encourages CJK users of both systems to consult both organizations and established uniformed matching criteria.

As the RLIN online system supports both the original and derived records, it offers the flexibility and latitude of creating the local versions. This is clearly an advantage.

CEAL should contact both organizations and establish the matching criteria for East Asian language materials. It is to OCLC and RLIN’s advantage to supply the matching criteria in order to maintain the integrity of their databases.

3. OCLC CJK’s hardware performance needs upgrading.

OCLC CJK uses the 8-bit micro computer, the IBM PC, which is slow for a large and complicated program like CJK (a total of 4 megabites, as compared with the M300 software* which contains only 0.25 mega-bites.) OCLC should upgrade its
CJK to the 16- or 32-bit hardware as soon as possible. Generally, the AT and its compatibles can run the PC and its compatible software without modification. (The reverse is not true.) However, because of one constraint with the current OCLC CJK hardware, the software must be modified in order to go through the higher clock speed of the AT level configuration.

In 1984 when OCLC started to develop its CJK, the AT cost substantially more than a PC. It made sense to consider the PC at that time. By 1986 when the CJK was introduced, the AT, (which was discontinued by then,) or its compatibles cost less than the PC in 1984. The AT-compatible prices have dropped down much further in the last two years. The continued choice of the PC for OCLC CJK has become a concession to its users.

4. OCLC CJK's vernacular capability is easy to learn yet the character entry procedure is somehow cumbersome.

RLIN incorporated all CJK characters into one character set. OCLC inherited this character set to make record exchange easy. In its CJK character entry programming, OCLC divides this character set into subsets to improve the operating efficiency. This concept, though logical in theory, turns out to be a barrier. Because the ideographic characters (Chinese full and simplified, kanji, and hancha) simply can't be divided definitively, because some characters are used in more than one scripts and some are linguistically controversial in terms of script types. As a result, users have to switch between the character subsets frequently. Each time when the users switches to another subset, they have to wait for the new software loading procedure. This procedure creates additional keystrokes and breaks the consistency of typing. Meanwhile, because the speed of the PC is slow, the users are forced to seat and wait for the new character table loading procedure.

To solve this problem, OCLC can consider a conditional branching subroutine to nest the subsets internally. At the end of each subset, e.g. the Chinese full character, when the desired
character is not found, the branching (e.g. the IF NOT . . . GO TO commands) will continue to search the next likely subset, e.g. the Chinese simplified, the Japanese kanji, and the Korean hancha. This user-transparent subroutine requires no change of the current keyboard operating procedure, yet reduces the need of toggling back and forth all the time.

5. OCLC merges the CJK records with all other records, as a result, each search key has to go through the entire database currently containing more than 18 million records. To increase the speed of the existing search structure, OCLC can add an indexing subroutine to sort out the CJK records. Each search key transmitted from a CJK workstation can be identified by the terminal I.D. or the escape code designating the CJK vernacular character set, and directly routed to the CJK records.

To do so, OCLC can (and should) index the language code in the fixed field and the 042 field. This additional indexing allows CJK search keys to go through the CJK records and bypass the other 17.75 records. Implementing this indexing subroutine requires minimal effort, yet saves CJK users tremendous searching time.

6. Search keys containing common terms, e.g. Chung-kuo (China,) or Nihon (Japan), or Koria (Korea,) are very inefficient because there are too many matching records. Following the stop word list for the roman language records, OCLC should develop and validate a CJK stop word list. The stop word list ignores the common terms and indexes the next significant words.

7. The OCLC online “traffic jam” forces the users to seat and wait too often. OCLC should develop a CJK database subset and maintain a separate telecommunication line for the CJK users. This subset can be easily established with the same language code indexing subroutine and requires no physical separation. There will be no impact on the romanized CJK search keys entered from
non-CJK workstations. The development cost for this separate communication line is minimal. To establish and maintain a separate this line for CJK, OCLC may incur additional costs of logistics that have to be justified.

8. The OCLC users interface with the current online system on a field-by-field basis. This interface creates an overflow of traffic to be regulated. The OCLC New System is expected to reduced this congestion, as it interface on a record-by-record basis.

9. The records from research and academic libraries frequently contain out of ordinary, if not rare, characters. As the OCLC CJK does not support an online thesaurus, entering these characters is time-consuming and, at times, disruptive.

10. In the U.S., the standard of transliteration for Chinese, the Wade-Giles system, is based on *A Chinese-English Dictionary* (by Herbert Giles.) The OCLC CJK, however, uses *Mathew's Chinese English Dictionary* (by R.H. Mathew) for the Chinese character entry. This discrepancy confuses the users and creates a potential bibliographic data quality control problem.

   Overall, as compared with the manual cataloging, ASIA has benefited from the OCLC CJK system. After all, sharing cataloging reduces the rate of costly original cataloging. This is an overwhelming advantage of library automation.

V. A NEW APPROACH TO THE MARC FORMAT

The final general constraint on ASIA's productivity is the rigid structure of the MARC formats. MARC is the single most important and successful investment of the cataloging community. Standardized machine readable formats make possible large scale sharing of cataloging effort and access. The structure of MARC, however, is complicated. While it has remained the standard for
the past 20 years, the new concept of data processing and computer technology have progressed and opened several more efficient options. Among the new techniques, the relational database structure, random access technique, and advanced search and indexing algorithm in particular have offered the library community simpler and lower cost options.

As the general computer and information industry is striving for simpler, and more flexible programming languages similar to human languages, the library community too have to modify, if not entirely redesign the MARC. Otherwise, our investment on MARC may reach the point of diminishing return. It does not only take students, but also experience catalogers a long learning curve to acquire the skills of the MARC cataloging. Moreover, the rigid field and subfield structures of MARC further slow down their productivity even after they have acquired the skills, because of the need to consult many brick-size manuals and to type field tags and subfield delimiters. While most data elements of the leader and fixed field of MARC, such as the language code, bibliographic and encoding levels provide advantages for library data processing, the 999 logical fields each containing various number of subfields are excessively complicated.

Perhaps, the MARC structure would not have to be more complicated than using the 12 areas of the conventional catalog card device that has been so simple and natural to catalogers as well as the patrons. Several common affordable database management software for personal computers, e.g., dBASE, can adequately handle the proposed bibliographic data structure. (For a specialized application like MARC, the specific software, of course, has to be developed in order to efficiently handle this structure.) The following examples illustrate the efficiency and flexibility of this new database structure.

Taking the “M/F/B” of the fixed field for example, while I can understand the advantage of sorting out fiction, biography and non-fiction from one another, I can’t imagine the need to sort out the main entry and body of entry relationship (noted by
“M”.) In the rare case that this relationship must be sorted, then a common, simple search and indexing scheme can serve this purpose. If we create a data field called “M” for main entry, “TI” for title, “AU” for author, and “SR” for statement of responsibility. To sort the main entry and the body of entry relationship, a user only has to enter a simple search command, e.g. FIND or SEARCH or SEEK, and then qualify this command, e.g. FOR M = TI or AU or SR. The program will compare the main entry of each record against the elements of the body and list or count each record when there is a match.

The main entry corresponds to the 2nd area of the conventional catalog card. In the dBASE type of structure, a user can create a field call MAIN ENTRY. The title (and its variations, i.e. subtitles or parallel titles,) statement of responsibilities, edition, the publishers, etc. represent the succeeding fields or subfields. The fields and subfields can easily be identified by mnemonic names. For instance, the user can create a data field called “TI” for title. The title variations will become the subfields, e.g. “PT” for parallel titles and “ST” for subtitles. Better still, once the areas are assigned for the MARC data elements, many elements do not even need mnemonic names, because the ISBD punctuation marks automatically identify the fields or subfields. For instance, in area 4, the slash “/” automatically flags the statement of responsibility; in area 12, the arabic and roman numerals flag the subject headings and title tracings. Therefore, there is no need for additional tags.

Taking a couple of variable fields for another example, the uniform title has several variations tagged under 130, 240, and 243, and series are tagged under 490 or 440 according to whether it is traced or traced differently and its status in the name authority file. Indicators of each these fields further designate data processing functions. Actually, one physical field may be sufficient to manipulate the variations of the uniform title, or series statements. One field identified by different mnemonic names, e.g., MUT for the uniform title used as the
main entry and CUT for collective uniform title, is sufficient for the creating, storing, sorting, retrieving, and processing of these bibliographic data elements in different output formats. Similarly, the variations of series statements only require one physical field identified by different mnemonic names.

In other aspects, as the computer storage space becomes larger and cheaper and the searching speed gets faster, the name authority file and the bibliographic record can be linked instantly. This "normalization" eliminates the need of additional notation to flag the variations, such as 440 or 490. We no longer have to torture our memory and fingers to tag and type so many fields and subfields. The capabilities and flexibility of today's computer have offered us many simpler options of standardized machine-readable format for the bibliographic data.

As a matter of fact the computer technology and indexing technique can also make cataloging rules more flexible and easier to apply. For instance, once an authority file is established and maintained for a personal name or title, local (or national) practices bear no impact on the standards. Each library can use its own preference in creating and retrieving a record locally. For instance, because the normalization structure links each applicable field with the authority file. For instance, as long as one of the variations of Sir Winston Churchill is established in the name authority file as a common standard, libraries don't have to waste time finding out which form to use, as the linkage relationship allows the local variation to co-exist with or be differentiated from the national standard and to sorted before retrieval. As a result, catalogers don't have to spend a great deal of time in verifying name authority file. This flexibility and tolerance can also reduce the much frustration of the situation frequently described as "when in doubt" in the AACR rules. The power and functions of computers can also make cataloging easier in some other areas. For instance, memorizing or looking up the precedence of notes can be time-consuming. Computers can be programmed to check and maintain the precedence all the time.
regardless the inputting or numerical sequence. The cataloger’s mind can, therefore, be freed for more significant cataloging functions as the high priest of human knowledge.

My suggestion for the current MARC is that when a simple 9- to 12-field structure using mnemonic terms are adequate, there is little need to bring in 999 fields and a string of subfields within each field. Neither patrons nor computers has to know the concept of main entry or tracings or uniform titles. It is the cataloger who recognizes and interprets the bibliographic significance of these elements. Catalogers don’t need 100, 110, 111, 240, 245 etc. and their associated subfields — which at times, can be mind-boggling — to interpret these bibliographic elements. In terms of data processing functions, the Boolean logic relations between data elements (AND, OR, NOR, and XOR) can be readily applied for indexing, sorting, and retrieving bibliographic data down to any single character.

By the same token, using indicators to flag the non-filing characters serves little purpose in view of today’s indexing and searching techniques. The articles, “A” and “THE” can be indexed with a EITHER OR structure. That is they will be read and compared if included in the search key; and ignored, if absent. Excessively rigid structure not only slows down performance but also creates potential errors. Wrong tags — easy mistakes — make records “unsearchable,” “unprocessable,” or cause misinterpretation in data analysis.

Using numerical tags to flag the status of tracings creates another redundant effort. Again, the presence of the elements itself indicates the availability of tracing. Further, based on the ISBD notation, the arabic and roman numerals automatically designate the unique data identities of the tracings. Catalogers can recognize the presence of the tracings without the current MARC field numbers which at best, are arbitrary designations. For data processing, computers can also identify the tracings with the presence or absence of these fields without additional tags. Therefore, between men and machine, it is possible to have
a good relationship with no strings attached.

Speaking without the benefit of a comprehensive study, the dBASE type of structure and the 12 areas of the conventional catalog card layout emanate a common sense appeal. They do warrant further studies. MARC was designed 20 years ago, today, this structure may be too heavy a burden for libraries to carry. Most libraries are either planning or implementing automation. It may be wise to re-assess the available options before we entrust all our resources to one.

After all, we use machine to make our jobs easier, not to complicate them. MARC cataloging may not have to be more difficult than the natural flow of human intellectuality and deft. In manual cataloging, we write on a piece of paper; in MARC cataloging we type on the screen. It can be done this way.

**FOOTNOTE**

* M 300 is the base from which the OCLC CJK extends.*