Quality Control of Chinese Monographic Records: A Case Study

Fung-yin K. Simpson

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

BYU ScholarsArchive Citation


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.
QUALITY CONTROL OF CHINESE MONOGRAPHIC RECORDS: A CASE STUDY

Fung-yin K. Simpson University of Illinois

Purpose of the study:

East Asian cataloging productivity in the U. S. has benefitted from many developments in the RLIN CJK and OCLC CJK systems since the 1980s. Recently developed information technology, such as the MS-Windows on-line environment, and the concept of resource sharing in cataloging between RLIN CJK and OCLC CJK, have resulted in great efficiency in CJK cataloging. After all of these developments, what is the current state of Chinese cataloging records? In order to answer this question, this study first reviews some of the recent developments in cataloging Chinese materials. I then go on to explore the quality of Chinese monographic records by checking critical errors which affect retrieval, and by comparing error rates among cataloging records from various sources: Library of Congress (LC); LC adapted records done by other institutions (LC-Copy); OCLC member institutions (Member); and OCLC tape-loaded from RLIN member institutions (Member-RLIN).

Recent Developments in Chinese cataloging

Developments in information technology

In the 1980's, RLG and later OCLC developed CJK (Chinese Japanese Korean) systems that have allowed catalogers to display and input CJK scripts in the bibliographic databases. In the 1990's, resource sharing was established between RLG and OCLC to allow their CJK records to be exchanged. Moreover, both RLG and OCLC have been seeking international cooperation to augment their bibliographic databases and to expand their functionalities in searching, inputting, and displaying their CJK records. Two examples of their accomplishments will suffice. RLG incorporated five libraries from China and thirteen RLG-member libraries to create an international union catalog of Chinese rare books with imprint years range from 1080 to 1795. OCLC tape-loaded nearly 300,000 Japanese records from the Waseda University Information Network (WINE) into the database in December 1995.

Both RLG and OCLC have developed continuing technical improvements to their systems. The OCLC CJK Plus system was introduced in May 1993 with MS-Windows features. This development made it much easier to edit CJK bibliographic records such as using copy and paste functions. Another convenience is the availability of multiple screens to flip through several bibliographic records. After November 1994, the CJK Plus system migrated to the PRISM service, and the searching capability expanded to include phrase search, Boolean combined keyword searches, and other useful searches. A program for Pinyin to Wade-Giles romanization conversion as well as record validation became available at this stage. In 1995 RLG released...
software compatible with MS-Windows. In the following year, inputting methods for CJK characters were expanded to include Pinyin and Wade-Giles romanizations, along with the character components inputting method. All the above functionalities have advanced CJK cataloging into a more user-friendly and technologically advanced cataloging environment.

Resource sharing in CJK cataloging:

RLG and OCLC have tape-loaded each other's member library's CJK monographic records since November 1989. However, more editing is needed to clean up the tape-loaded records in order to meet the inputting standards for each system. One example involves the use of word division and aggregation. In RLIN CJK, a special symbol called an aggregator is used to join individual Chinese characters together to form a semantic unit. Tape-loaded Chinese records from OCLC do not have aggregators, but they are required by RLG. On the other hand, vernacular punctuation and spaces in RLIN are not acceptable to OCLC. Furthermore, the tape-loaded records in both databases usually lack call numbers.

OCLC-member libraries have been enhancing the tape-loaded records with appropriate call numbers. Since December 1993, OCLC has tried to retain the call number in the RLIN-CJK records that are tape-loaded into the OCLC database, but only a part of the selected RLG members' records were implemented initially. The number of records lacking a call number among the tape-loaded records is decreasing, but constant editing for extra space and proper punctuation is still required.

LC copy cataloging:

In the Fall of 1993, the Library of Congress started to treat copy cataloging as a standard activity in order to increase cataloging output and reduce its arrearage. External source records from bibliographic utilities, such as OCLC, RLIN and other sources, have been adapted for this purpose. The records done as copy cataloging can be identified by "lccopycat" in the 042 field. Because these records have been re-examined and approved by LC, this development should increase the accuracy of the records in the databases.

Methodology:

Although this is a small case study done in a local institution, the statistical analysis should provide some important observations concerning quality control of Chinese bibliographic records. The data consisted of 380 Chinese monographic records selected from OCLC's WorldCat (the OCLC Online Union Catalog) which were processed between October 1995 and February 1996 in the Asian Library of the University of Illinois at Urbana-Champaign. All monographic records processed during this period were considered. Although this is not a random sample of cataloging record in the OCLC database, it is reasonable to assume that the records processed during this period are typical rather than exceptional. The advantage of this approach is that each item is physically available for verification, along with the matching bibliographic record.
In order to limit the scope to current publications and to avoid retrospectively converted records, reprinted titles and titles published before 1987 were excluded, as were bibliographic records entered before 1990. The selected titles included 31 titles published in the late 80's and 349 titles published from 1990 to 1995. For each bibliographic record, the following fields were examined: the date entered in the WorldCat, the imprint date, the presence of the LC Classification Number, the presence of LC Subject Headings, and the cataloging institutions. These fields as well as the numbers of errors were entered into a Microsoft EXCEL spread sheet. Table 1 shows 10 records to illustrate the structure of the data. The data from the spread sheet were exported to SAS-JMP for statistical analysis.

**TABLE 1**

LIST OF ELEMENTS IN THE BIBLIOGRAPHIC RECORDS

<table>
<thead>
<tr>
<th>Record</th>
<th>E-yr</th>
<th>P-yr</th>
<th>LCCN</th>
<th>LCSH</th>
<th>Institution</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94</td>
<td>93</td>
<td>y</td>
<td>y</td>
<td>member</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>95</td>
<td>93</td>
<td>y</td>
<td>y</td>
<td>Lc-copy</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>95</td>
<td>93</td>
<td>y</td>
<td>y</td>
<td>Lc</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>95</td>
<td>94</td>
<td>y</td>
<td>y</td>
<td>member</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>95</td>
<td>93</td>
<td>y</td>
<td>y</td>
<td>member</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>95</td>
<td>94</td>
<td>y</td>
<td>y</td>
<td>member</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>95</td>
<td>94</td>
<td>y</td>
<td>y</td>
<td>Lc-copy</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>94</td>
<td>94</td>
<td>y</td>
<td>fiction</td>
<td>m-rlin</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>93</td>
<td>92</td>
<td>y</td>
<td>fiction</td>
<td>member</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>90</td>
<td>87</td>
<td>y</td>
<td>y</td>
<td>m-rlin</td>
<td>1</td>
</tr>
</tbody>
</table>

For completeness of the bibliographic records, in addition to checking the existence of the LC Classification Number and LC Subject Headings, error counts were constructed by checking the crucial errors likely to affect the retrieval of the record. Examples include mis-romanization, tag and code errors, incomplete and omitted fields. The correlation between error rates and the cataloging institutions were determined and compared.

Results and Analysis

Among the 380 sampled records, 89 records were cataloged by LC (23.4%), 59 records were cataloged through LC-Copy (15.5%), 152 records were cataloged by OCLC member institutions (40%), 80 records were cataloged by RLG member institutions and tape-loaded into OCLC (21.1%); see Table 2. The percentage distribution in this study suggests that LC contributed roughly a quarter of the Chinese records. The remaining three quarters were contributed by member institutions either from OCLC or from RLG.
TABLE 2
PERCENTAGE OF CATALOGING INSTITUTION

<table>
<thead>
<tr>
<th>Institutions</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>89</td>
<td>23.4%</td>
</tr>
<tr>
<td>LC-Copy</td>
<td>59</td>
<td>15.5%</td>
</tr>
<tr>
<td>Member-OCLC</td>
<td>152</td>
<td>40.0%</td>
</tr>
<tr>
<td>Mem.-RLIN</td>
<td>80</td>
<td>21.1%</td>
</tr>
<tr>
<td>(Tape-loaded)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>380</td>
<td>100%</td>
</tr>
</tbody>
</table>

Completeness of the Chinese records

The provision of LC Classification numbers and LC Subject Headings, and adding parallel Chinese scripts to the core fields are the three basic criteria to evaluate the completeness of the Chinese records. From these 380 records, 43 records lacked LC Classification numbers. Among them, 16 records without LC classification numbers were the result of inappropriate tape-loading (see Table 3). Six records did not assign any LC Subject Headings, excluding fiction and other literary works. Six romanization-only records lacked parallel Chinese scripts, which are required for standard CJK cataloging. In this study, approximately 14.5% of Chinese bibliographic records are incomplete with respect to the basic requirements of CJK bibliographic records. Almost all of these incomplete records were provided by the member institutions, except one LC Minimal level (level: 7) record without an LC classification number.

TABLE 3
LC CLASSIFICATION NUMBER

<table>
<thead>
<tr>
<th>Institution</th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>1</td>
<td>85</td>
<td>86</td>
</tr>
<tr>
<td>LC-Copy</td>
<td>0</td>
<td>62</td>
<td>62</td>
</tr>
<tr>
<td>Member (OCLC)</td>
<td>26</td>
<td>126</td>
<td>152</td>
</tr>
<tr>
<td>Mem.-RLIN</td>
<td>16</td>
<td>64</td>
<td>80</td>
</tr>
<tr>
<td>(Tape-loaded)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>43</td>
<td>337</td>
<td>380</td>
</tr>
</tbody>
</table>
Error counts and error rates

In addition to the completeness of the bibliographic records, the error rate is another vital measurement for the quality control of the bibliographic record. Each record was carefully investigated in fixed fields and variable fields. The checking elements were based on the requirements of Program for Cooperative Cataloging (PCC) Non-Roman core records. The most common errors were categorized into five major groups: code errors, rule errors, misspelling, ISBN errors, and additions for additional entries (see Table 4).

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Code Err.</th>
<th>Rule Err.</th>
<th>Misspell.</th>
<th>ISBD Err.</th>
<th>Additions</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>fix:enc</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>fix:cont</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>fix:ill</td>
<td></td>
<td></td>
<td>9</td>
<td></td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>fix:dates</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>fix:others</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>020:isbn</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>04x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>09x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1xx:Rom.</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>1xx:Ver.</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>245:Rom.</td>
<td>5</td>
<td>14</td>
<td>23</td>
<td>13</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>245:Ver.</td>
<td>5</td>
<td>12</td>
<td>4</td>
<td>3</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>246:FI</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>250:Rom.</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>250:Ver.</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>260:Rom.</td>
<td></td>
<td>12</td>
<td>21</td>
<td>8</td>
<td>0</td>
<td>41</td>
</tr>
<tr>
<td>260:Ver.</td>
<td>1</td>
<td>12</td>
<td>5</td>
<td>0</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>3xx</td>
<td>1</td>
<td>4</td>
<td>32</td>
<td>7</td>
<td>14</td>
<td>58</td>
</tr>
<tr>
<td>4xx&amp;8xx:R</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>4xx&amp;8xx:V</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5xx</td>
<td>6</td>
<td></td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>6xx</td>
<td>6</td>
<td>17</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>38</td>
</tr>
<tr>
<td>7xx:Rom.</td>
<td>1</td>
<td>10</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>7xx:Ver.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>total</td>
<td>62</td>
<td>108</td>
<td>114</td>
<td>75</td>
<td>71</td>
<td>430</td>
</tr>
<tr>
<td>%</td>
<td>14.4%</td>
<td>25.1%</td>
<td>26.5%</td>
<td>17.4%</td>
<td>16.5%</td>
<td>100%</td>
</tr>
</tbody>
</table>

35
Code errors are incorrect or omitted MARC coding for tags, indicators and subfield codes. Failure to encode a fixed field required for the bibliographic input standards is counted as an additions. For tape-loaded records, if the LC Classification number was retained or enhanced, but the encoded level remained as L (less than full), then it did not count as a code error. Contraventions of AACR2 and LC rule interpretations (LCRIs) are categorized as rule errors. Incomplete/missing transcription of the title or publication information are also treated as rule errors. Contradictions with the LC authority file in name entries are also counted in this category.

In addition to the regular English typographical and misspelling errors, there are several other types of misspelling errors in Chinese records: mis-romanization, or incorrect Wade-Giles romanization in letters or diacritical symbols, improper use or omission of hyphens for personal or geographic names, and wrong Chinese characters in vernacular fields. Extra spaces between romanized words and vernacular words were not counted as ISBN errors. Due to the different cataloging standards of OCLC and RLG, the tape-loaded records from RLG have not been in accordance with OCLC’s requirements for CJK (Chinese Japanese Korean) cataloging, so unnecessary spaces and word separations frequently appear in the records. This situation also occurred in many of the LC records which were cataloged through RLIN.

Frequent error occurrences

From Table 4 it can be seen that Field 245 (Title statement) and Field 260 (Imprint information) have high error rates. This finding is similar to Lei Zeng’s research for both OCLC and RLIN samples. However, Zeng indicated that inadequate ISBN punctuation and spacing, missing entry, and inadequate space in editing were the three most frequent errors, followed by misspelling of romanized words. According to this study, misspelling in romanized fields is the most frequent error, followed by rule errors, then ISBN errors in both fields. Since 1994, when the OCLC CJK system was implemented in the PRISM environment, the problem of improper spacing was drastically reduced for full record-basis editing, and some of the tagging errors also can be detected and corrected through record validation. Therefore, the inadequate spacing problems mentioned in Zeng’s research have been reduced significantly over the years.

In the current study, misspelling in the romanized fields was the most frequent error. It has justified the controversy and unpopularity of the Chinese romanization system. Even a native speaker has to take time to learn how to romanize accurately enough to search the database. Inaccurate Chinese characters rarely occurred, but some regularly used characters cannot be found in OCLC CJK character input codes.

Rule errors in Fields 245 and 260 consist mostly of improper or missing transcription of title information or inconsistency between romanized and vernacular fields. Zeng pointed out the same error occurrences in her study. It is common to have a complex description on the Chinese title page with various forms of title headings and authorship, the publication information on the colophon page often states the publisher as well as distributor with long official title names. Furthermore, the required parallel fields with roman and vernacular have apparently exhausted the patience of Chinese catalogers, so missing entries or inconsistency in 245 and 260 are common. This is apparently due to human error while inputting the record, rather than unfamiliarity with AACR2 or LCRIs.
Another common rule error is the name entry conflict with the LC authority entry in fields 1xx, 6xx, and 7xx. Many CJK personal names share the same romanizations and CJK authors often use various forms of headings in publications. However, LC's authority file only provides the roman form for CJK personal names, and it cannot effectively distinguish one name from another. Consequently, this has often confused catalogers in their choices of the name entries for particular Chinese personal names. Some name entries in the records were added with other types of romanization which conflict with LC's authority entries. These should be added locally to adjust for the need of patrons, if it is necessary.

Other kinds of errors are not critical, but they affect nonetheless affect retrieval. In particular, additional name entries and subject entries are important. As I mentioned earlier, only six out of 380 records lacked subject headings, excluding fiction and other literary works. However, catalogers should take care to note when more subject headings could be provided in order to convey complex subject matter, and provide added name entries to indicate more than one authorship or editorship. It is important to record the ISBN, if known, accurately, because the ISBN is one of the important key elements for searching foreign publications in the databases, in case the searcher is not familiar with the language, or the system of romanization which has been used in the database.

Error rates among cataloging institutions

There were 430 errors identified in the descriptive parts of the 374 records. Six records with only romanization and lack of parallel Chinese fields were counted as maximum of 10 errors for each. In all, there were 490 errors for 380 records in descriptive cataloging. The mean error rate per record was 1.29.

Compared among the cataloging institutions, LC had the smallest error rate with 0.80 (+0.12), followed by LC's copy cataloging with 0.864 (+0.16) errors, OCLC had 1.71 (+0.19) errors, tape-loaded records from RLIN had 1.35 (+0.21) errors (see Table 5).

In addition to the above errors in descriptive cataloging and subject headings, there were 43 records lacking of the LC Classification Numbers: 26 OCLC records, 16 tape-loaded records, and one LC Minimal Level (level 7) record. Consequently, the mean error per record is 1.40. LC still maintained the least error rate with 0.81, OCLC had 1.89 errors, and tape-loaded records had 1.55 errors per record. In order to test significance of all the difference between the means, the Tukey-Kramer Honestly Significant Difference (HSD) method was used with an overall significance level of 5 %. The Tukey-Kramer analysis shows that the member library error rates (OCLC and tape-loaded) were significantly higher than LC rates (original and copy cataloging).
### TABLE 5
**ERROR RATE BY CATALOGING INSTITUTIONS**

<table>
<thead>
<tr>
<th>Institution</th>
<th>Number</th>
<th>Mean</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>89</td>
<td>0.79775</td>
<td>0.12293</td>
</tr>
<tr>
<td>LC-Copy</td>
<td>59</td>
<td>0.86441</td>
<td>0.15937</td>
</tr>
<tr>
<td>Member (OCLC)</td>
<td>152</td>
<td>1.71503</td>
<td>0.19461</td>
</tr>
<tr>
<td>Mem.-RLIN (Tape-loaded)</td>
<td>80</td>
<td>1.35000</td>
<td>0.20905</td>
</tr>
</tbody>
</table>

Mean Error Rate Per Record: 1.289474

As for the member institutions, the result does not indicate that RLIN records have fewer errors than OCLC. The required input standards and composed members in these two cataloging utilities are different. The six romanized-only records included in this study were cataloged by OCLC member institutions that may not have CJK workstations available for their cataloging needs. But all of the tape-loaded records from RLIN into OCLC were pre-selected from RLIN member libraries, at least in this case. For instance, by counting records paired with Chinese scripts and excluding the six romanized-only record, OCLC had 1.37 mean errors in descriptive cataloging, which is very close to the error rate of 1.35 for tape-loaded records.

**Conclusions**

The baseline study that was done by Leazer and Rohdy in 1993, showed that LC produced 91.2% of the bibliographic records for domestic publications, 61.8% of records for Great Britain publications, and 38.2% of records for other foreign publications. In this study, 23.4% of Chinese bibliographic records were cataloged by LC, and it maintains the high accuracy with 0.81 errors per record. Member institutions cataloged more than 75% of Chinese monographs in the database. Because of resource sharing in CJK cataloging, exchanging bibliographic records between OCLC and RLIN not only expanded both databases, but also increased the availability of CJK bibliographic records in a timely manner to their member institutions.

In the foreseeable future, OCLC and RLIN will be likely to convert more foreign MARC records into their cataloging databases. The conversions no doubt will provide more timely access for foreign bibliographic resources in North America. It will benefit more in collection development and reference, but catalogers may have to spend more time to enhance these records in order to maintain high quality control in the databases.

With the complexity of the Chinese records, it has been difficult to maintain quality control. The above study in error rates shows that the most frequent type of error (26.5% of all errors—see table 4) is mis-romanization in various fields. It is widely recognized that romanization cannot properly differentiate thousands of unique Chinese characters effectively. However, it nonetheless remains a common method to integrate Chinese records into a unified online catalog environment in North America. Currently, the online public access catalogs of most academic libraries do not have CJK displays. Regardless of its flaws, romanization is the system of access
currently in use. It should therefore be standardized, and it should have clear guidelines for catalogers and library patrons to follow.

The second most common category of errors is rule errors, which made up 25.1% of all errors calculated in this study (see table 4). Incomplete or omitted transcriptions in the entries are the most frequent errors of this type. Compared to regular English bibliographic records, Chinese bibliographic record require a lengthy and complex process to create parallel fields in romanization and Chinese scripts. It is important to apply the concept of simplification in cataloging Chinese materials, especially for Field 245 and Field 260.

Even more than this, it is crucial to have a comprehensive CJK cataloging manual which incorporates rules and special guidance for CJK catalogers to follow. At present, the only AACR2 Workbook for East Asian publications was published in 1983,12 and it needs to be revised to adhere to the current rule changes for the current cataloging environment. Last but not least, more research should be conducted in bibliographic control of CJK materials, so that complicated cataloging issues can be addressed and standardization in cataloging can be achieved.

NOTES


10. Sall, John and Ann Lehman. JMP Start Statistics: a guide to statistics and data analysis
