This paper explores and analyzes the errors in the sampled catalog records of Chinese monographs in the OCLC database. The author examined the errors by content, frequency, type and MARC field location to study the effect on both copy cataloging activities and library patrons’ access to an online catalog record. Finally, a number of recommendations are proposed for future research and error decreasing measures.

Headings:

Copy cataloging of Chinese monographs

OCLC database record evaluation
AN EMPIRICAL STUDY ON OCLC CATALOG RECORD ERRORS FOR THE COPY CATALOGING OF CHINESE MONOGRAPHS

by

Bin Jia

A Master’s paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Information Science.

Chapel Hill, North Carolina
April 2007

Approved by

__________________________
Jerry D. Saye
Chapter I Introduction

To better serve patrons from diverse background and support global research, libraries in the United States are expanding their holdings in languages other than English. According to the Online Computer Library Center (OCLC) statistics, Chinese is among the top ten languages in WorldCat. The Chinese language-coded records have a fast growth rate (6.6%) from July 2005 to January 2006. An American Council of Learned Societies’ (ACLS) report also indicated that East Asian libraries in the US doubled their holdings each decade from 1930 to 1985.

The creation of the CJK (Chinese Japanese Korean) cataloging by the Research Libraries Group (RLG) and later by OCLC in the 1980s enabled the full input and display of CJK scripts in bibliographic databases. A series of technical developments in the 1990s added functionalities to enhance software compatibility and further support Chinese romanization display. This greatly facilitated a true and complete reflection of Chinese bibliographic records. However, despite the efforts made, the catalogers’ task of creating accurate Chinese cataloging records is still quite challenging since it not only requires familiarity with cataloging rules, but also knowledge of the Chinese language. Inaccurate data input tends to hinder catalog access for the patrons and result in needless failed searches.

The purpose of a library is to provide qualified service to patrons. Among its main units, the cataloging department plays an important role in ensuring the quality of all the catalog records. In addition to the creation of the catalog records, maintenance and
enhancement of the catalog needs to be done so that records can be accessed easily and accurately. Easiness means that patrons can quickly find the wanted records by entering the index entries for a work, such as title, author or known subject, etc. Further, the accurate description of a work is an essential element to ensure that the patrons are getting the exact book they want.

The accurate description and fast processing of new acquisitions also relates to work quality in other departments, such as circulation, collection development and reference. An item can only be accessed by patrons when a cataloger updates the book information in the library system to indicate the book’s availability. To avoid the possibility of duplicate holdings and make purchase decisions, a careful pre-search by the collection development needs to be done with an examination on the existing catalog records. A catalog’s incorrect reflection of holdings can leave a reference librarian in an embarrassing situation since he’s unable to provide satisfactory answers to patrons’ questions. The above situations definitely harm the efficient operation of a library. The efforts of various departments may be involved to have the catalog corrected.

With a large number of works being added to the library, how can catalogers ensure the quality of bibliographic records while maintaining processing efficiency? The development of the Machine-Readable Cataloging (MARC) format in the 1960’s and the advent of OCLC have significantly changed cataloging practice in modern libraries. Currently, there are more than 57,000 libraries around the world using the services provided by OCLC.

Dedicated to the resolution of fostering information access and reducing cost per record, in the 1990’s, OCLC has been seeking international cooperation to augment its
bibliographic databases and to expand its functionalities in searching, inputting, and displaying records (Beall, 2004). In Fall 1993, the Library of Congress started to treating copy cataloging as a standard activity in order to increase cataloging output and reduce its arrearage (Simpson, 1998). Copy cataloging means preparing a bibliographic record by using, or adapting, a bibliographic record already prepared by someone in another library or organization (Schultz, 1995). This resource sharing approach has resulted in great efficiency in English language as well as Chinese language.

Despite the benefits of eliminating duplication of effort by the collaborative work of the OCLC members, various types of errors have been found in the databases for Chinese bibliographic records which potentially hinder the catalog retrieval and the proper record display for the library users. The errors might originate from the lack of adherence to cataloging rules, inconsistent formatting, careless typos, the complicated characteristics of Chinese language and its romanization form, etc.

Various approaches have been adopted for the consistent and error-free presentation of Chinese bibliographic records to avoid the necessity of cleanup work later. Some libraries take advantage of the employment of well-trained Chinese catalogers; some libraries upload records after examining them at least twice and some count on electronic tools such as online dictionaries to verify dubious script input in the record. Besides the knowledge backgrounds and experiences of individuals, one of the tricky reasons for catalogers to overlook errors is that libraries differ in the amount of quality control they exercise during the copy cataloging process. Indeed some might even omit the process of examining the bibliographic record in the source database because of pressures from handling large acquisitions (Beall, 2004).
To further facilitate the process of copy cataloging of Chinese bibliographic records to enhance the cataloging efficiency, the paper tries to answer the following questions: What are the errors in the record of the shared OCLC database? What are their types? What is their MARC location? What is their frequency? What are the possible reasons for the existence of these errors? Are there any methods to decreasing these errors?

The answers to these questions will contribute to both cost-saving of human resources in libraries and time-saving in record examination for the people who are doing the copy cataloging so that they can concentrate on more complicated tasks. The findings might suggest good conventions to follow.

In this paper, only Chinese monograph catalog records were examined since monographs are the groundwork for the transition to other formats such as continuing resources, i.e., series, microfilm, etc. Those formats are more complicated and will not be considered in this study.

The examination on both vernacular and romanization entries were conducted. For the sampled OCLC bibliographic record, Main Entry-Personal Name (100), Title Statement (245), Publication, Distribution, Etc (260), Physical Description (300), Varying Form of Title (246), Edition Statement (250), Series Statement/Added Entry-Title (246), Series Statement/Added Entry–Title (440), Series Statement (490), General Note (500), Subject Added Entry-Personal Name (600), Subject Added Entry-Corporate Name (610), Subject Added Entry-Topical Term (650), Subject Added Entry-Geographic Name (651), Added Entry-Personal Name (700), Added Entry-Corporate Name (710) and Series Added Entry-Uniform Title (830) are covered. The fixed field, call number field
(such as 050, 090, 082 etc), Standard Number field, i.e., 020, and the proper assignment on subject heading are excluded in this paper. This study did extensive observation on the record structure, formatting, content designation, and the detailed content of MARC bibliographic record.
Chapter II Literature Review

The examination on OCLC database is no longer a new field of study. Many studies have been done on the issues of bibliographic record quality. Most of the literature is oriented on one or two particular aspects of the bibliographic record; a study of it from a general perspective is quite rare. However, the literature examines the specific bibliographic record aspects thoroughly, analyzes the cause of the problems and proposes feasible solutions.

This literature review covers both English and Chinese language catalogs to see if there are similarities or differences between the two, so that the solutions identified for English language records might be considered for Chinese records.

The review focuses on three aspects: error types, reasons for the existence of the error and methods for decreasing errors in the OCLC database.

2.1 Error types:

In Schoenung's doctoral dissertation, he did research on the quality of the member-input monographic errors in OCLC. He deemed deviations between a sample record and its corresponding LC copy as errors, since they failed to follow the OCLC input standards. He also pointed out that a large portion of the variations were not really errors incurred by the inputting library, but rather were the result of changes in the OCLC input standards over the years. The errors covered in his work included content
designators (tagging), typographical errors, the phrasing of note fields, and the small
details of punctuation, capitalization, use of abbreviations, etc.

Beall’s paper concentrated on the typographical errors in English catalog records.
According to his study, only 35.8 percent of the English typographical errors had been
corrected by five selected libraries. The result indicated that the efficiency of the copy
cataloging is not equal to the improvement on accuracy of a catalog record. In this paper,
the situation for Chinese monographic records might be even worse since both the
Chinese vernacular and the corresponding romanization entries were required to be added
into the record.

In Fung-yin K. Simpson’s paper published in 1998, the error types and rates of
Chinese monograph records were investigated in addition to the completeness of the
bibliographic records. The main errors were categorized into five major groups: code
errors, rule errors, misspelling, ISBN errors and additions for additional entries. Besides
the regular English typographical errors and misspelling, there are several other types of
spelling errors in Chinese records: misromanization, improper use or omission of
hyphens for personal or geographic names, and incorrect Chinese characters in vernacular
fields.

The issue of record consistency is also problematic in the Chinese material
cataloging in OCLC. According to Yue Li (2004), the contradictory nature of the record
input standards did cause confusion for the catalogers as well as the library users. In the
paper, she mentioned several inconsistency issues by examining whole bibliographic
records including the fixed fields. The issues are: inconsistency in romanizing Chinese
vernacular, incompatible spacing in Pinyin and vernacular fields, U.S. Dollars vs.
Chinese Yuan in field 020, incorrect language coding in field 041, faulty input of Chinese language scripts, and translation of non-Chinese characters in the parallel description fields.

Jie Huang's article talked about the problem of word division particularly. The author pointed out that catalogers may interpret and segment titles incorrectly, resulting in inconsistencies in cataloging. Furthermore, users may enter search terms in wrongly aggregated Pinyin. These inconsistencies in word division between cataloger generated records, and between these records and user-input queries, will arguably affect retrieval in a negative way. Huang is strongly in favor of the cataloging and retrieving approach from Peking University Library’s (PKUL) and hope it can be adopted by the library community at large.

2.2 Reasons of the errors existence

The reasons of the error existence are quite complicated, they might relate to something intangible such as cultural background, changes in the cataloging input standards, policies, MARC format and systems both from the related organizations and also the failure to immediately complete corrections or updates.

Clement (2001) talked about the adaptation of Chinese romanization system from Wade-Giles to pinyin. The author’s in-depth analysis of the pinyin system, especially in the North American environments provides the basic knowledge background for exploring possible solutions. In the paper, some culturally related literature reviews are also provided to explain the sources of the errors.
Efforts have been made to eliminate the errors in Chinese records. The intricate nature of Chinese script along with the extensive use of both the traditional and simplified Chinese characters plus the difficulty in determining if a surname is known are some of the reasons that lead to error existence. Hu’s paper (2000) did comprehensive research on the problems of cataloging Chinese names and he analyzed the origins of the errors.

Beall’s paper pointed out that due to the different standards of quality control in libraries, the ability to copy data from other libraries can potentially detract the value it adds to the cataloging process. Libraries that copy data from a bibliographic record in the source database can also copy the errors made in the record.

2.3 Methods of decreasing errors in the OCLC database

Although the articles mentioned above suggested some potential solutions to decreasing the errors, they are unlikely to provide directly employable solutions that might be proposed in this research for their emphasis on a few aspects, and the following literature provided possible supplements.

In Tamas E. Doszkocs’s paper, he pointed out that not everyone spells correctly, so an added feature or a special function that enhances spelling accuracy should be considered. He did not try to impose a rigid dictionary for people to follow or consult to, but instead, a spelling “suggestion” system that could possibly reduce the occurrence of spelling errors. Since the Chinese monograph input in OCLC is much more complicated, a new model needs to be developed based on Tamas’ idea.
Although Jian’s paper mainly talks about issues in cataloging Chinese series, the paper provided background introduction to Chinese book publishing history, convention and its development. According to Chu, book publishing in China is making its way towards matching the rest of world in regards to improved quality and better management. In the future, standardized publishing conventions might potentially be a factor in facilitating catalogers’ work in finding main source information for books, providing users with more access points.

There are measures recommended in Beall’s article regarding the elimination of typographical errors. First, libraries can search their catalogs for common typographical errors according to an error list. Second, utilities and other suppliers of bibliographic records can routinely search and correct errors in their master databases. Third, utilities need to increase the incentives for enhancing master records by correcting typos, and they should make it easier for libraries either to correct the typos or to report them to the utilities’ quality control department.

According to Schoenung, OCLC, regional networks and member institutions have jointly created a variety of quality control mechanisms to promote the contribution of quality cataloging over the years, the major OCLC quality control features are input standards, automatic error detection, error reports, update reports, and enhance capability. This necessitates that the errors be fixed quickly if the local cataloger can report them to OCLC in a timely manner.

Most libraries also do later cataloging maintenance for a more correct and complete record display. However, contrary to expectation, compared with the measures
taken at the time of a record’s creation, the later maintenance on it is much more time-consuming and labor-intensive.
Chapter III Methodology

Some existing literature provided sound background knowledge on designing and implementing the research. This paper will take advantage of the existing methods and make further developments for new findings.

The findings of this research will contribute to the following areas:

1. The data analysis may help the catalogers identify the current prevailing errors of Chinese monographic records.

2. The efficiency of copy cataloging can be further improved with a quick check of the most frequent error types and locations for library’s local use.

3. The suggested methods for decreasing the errors may foster the development of algorithms for error detection so that copy cataloging could be a less expensive process.

3.1 Data source

The data come from the East Asian Collection (EAC) department at Davis Library at the University of North Carolina at Chapel Hill (UNC-CH). It is worth mentioning that the Chinese book holdings at UNC are among the top in the Southeastern US. On June 30, 2004, the East Asian materials in the Academic Affairs Library (AAL) at UNC-Chapel Hill reached 131,230 volumes, of which 125,160 are in Chinese. The largest volumes of the materials (32%) are in Chinese language and literature.
From the most recent Annual Report of EAC for the academic year 2005-2006, there were about 2,000 Chinese titles cataloged for the collection. The following statistics provide an overview of the cataloging activities in the EAC and the total volume has constant growth around 6% each year:

<table>
<thead>
<tr>
<th>Month</th>
<th>Titles</th>
<th>Total Volumes</th>
</tr>
</thead>
<tbody>
<tr>
<td>July</td>
<td>59</td>
<td>263</td>
</tr>
<tr>
<td>August</td>
<td>167</td>
<td>220</td>
</tr>
<tr>
<td>September</td>
<td>173</td>
<td>279</td>
</tr>
<tr>
<td>October</td>
<td>222</td>
<td>275</td>
</tr>
<tr>
<td>November</td>
<td>193</td>
<td>370</td>
</tr>
<tr>
<td>December</td>
<td>190</td>
<td>200</td>
</tr>
<tr>
<td>January</td>
<td>196</td>
<td>361</td>
</tr>
<tr>
<td>February</td>
<td>179</td>
<td>348</td>
</tr>
<tr>
<td>March</td>
<td>200</td>
<td>230</td>
</tr>
<tr>
<td>April</td>
<td>127</td>
<td>382</td>
</tr>
<tr>
<td>May</td>
<td>107</td>
<td>218</td>
</tr>
<tr>
<td>June</td>
<td>168</td>
<td>262</td>
</tr>
<tr>
<td>Total</td>
<td>2,101</td>
<td>3,408</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>By Format</th>
<th>Chinese</th>
<th>Total Titles</th>
<th>Total Volumes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monographs</td>
<td>1,980</td>
<td>2,643</td>
<td></td>
</tr>
<tr>
<td>Serials</td>
<td>42</td>
<td>451</td>
<td></td>
</tr>
<tr>
<td>Microforms</td>
<td>2</td>
<td>195</td>
<td></td>
</tr>
<tr>
<td>Films (VCD, DVD, VHS)</td>
<td>72</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>CD-ROM</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Accompanying CD</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Electronic Resources</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2,101</td>
<td>3,408</td>
<td></td>
</tr>
</tbody>
</table>

With the large acquisition of Chinese monographs, copy cataloging is the primary cataloging activity in the East Asian Collection and it contributes to the cataloging efficiency. It is the department’s policy to have the matching bibliographic record from
OCLC WorldCat printed and checked by at least two student assistants from the School of Information and Library Science at UNC-CH. The EAC cataloging specialist is responsible for monitoring the process and conducting a final check before uploading the record to the library’s online catalog. The students mainly check the descriptive record. The cataloger is responsible for the fixed field and the call number in addition to rechecking the students’ work. This process was developed based on past experiences and was designed to ensure record quality.

3.2 Sampling

A sample of the OCLC records used in the EAC is the core data in this study. Considering the study is proposed for a master’s paper and the total available time is only three to four months, an analysis of a large data set was not feasible. Since the data need to be checked and corrected, a comprehensive retrieval and examination in WorldCat also is not possible. It is felt that the most recent one to two months’ cataloging production is sufficient size to draw a useful sample.

The data need to be sampled from the EAC at the Davis library at UNC-CH for the following reasons:

1. Accessibility

   The record data can be easily accessed through the OCLC Connexion in the library and the printed bibliographic records bundled for storage monthly

2. Correctness

   The tenet of the cataloging in EAC is to provide accurate catalog records and more access points for its patrons. During their examination of the record,
both the student assistants and the cataloger are responsible for the identification of errors. Whenever there’s any confusion for the students, assistance from the experienced cataloger provides guidance.

3. The representative processing workload

Sampling was based on the normal workload of the EAC (over 200 volumes of monographs a month), so the total number of records examined is in the range of 400-500. The sampled data had the percentage of errors and other statistical data recorded.

3.3 Data measurement criteria

In this study, OCLC record quality measurement criteria are based upon the Descriptive Cataloging of East Asian Material: CJK Examples of AACR2 and Library of Congress Rule Interpretations, the Anglo-American Cataloguing Rules, Second Edition (AACR2), the AACR Workbook for East Asian Publications, Cooperative Online Serials (CONSER) Cataloging Manual, 2002 Edition and the LC MARC standards.

LC’s Chinese Romanization Policies for Cataloging Chinese Material and Romanization of Chinese Geographic Names in Descriptive and Subject Headings are references for checking Chinese romanization entries. For Chinese scripts, references were made to both authoritative Chinese dictionaries and some online dictionaries.

3.4 Analytical techniques
The raw data were recorded when an error was identified when doing copy cataloging for Chinese monographs. Each error was recorded under the predefined name of the error record field.

We have the error data recorded horizontally in the following data collection table:

<table>
<thead>
<tr>
<th>Monograph #</th>
<th>M Field #</th>
<th>Error 1</th>
<th>Error 2</th>
<th>E type 1</th>
<th>E type 2</th>
<th>246 or 440</th>
</tr>
</thead>
</table>

For the complete record of the errors, the table was extended horizontally with the following fields:

1. **Monograph number**
   Unique numeric number was assigned to each monographic record in an ascending order. The number not only helped the counting of the total number of monographs but also the quick access for reexamination on the detected errors. The same number can be repeated several times in rows under the monograph number field if errors were detected in more than one MARC fields of each record

2. **MARC field number**
   If there are any errors identified in a bibliographic record, the corresponding MARC field tag, e.g., 100, 245, etc., were recorded

3. **Error 1 and Error 2**
   These two fields were designed for recording the detailed description of an error. It is quite possible to have more than one error appeared in a MARC
field. Although more places were reserved for the error description, they were
deleted later in this study because there were only at most two errors in a
single MARC field identified. Examples of the description of an error could
be: incorrect Chinese script, no edition information in the 250 field
4. Error type 1 and Error type 2
The corresponding error types were recorded for Error 1 and Error 2. The
example error types are: incorrect description, spacing error, formatting error,
etc
5. 246 or 440 field
The field was designed to record errors in the 246 and 440 fields

3.5 Final data set
Processed OCLC bibliographic records in the month of November and December
of 2006 in the EAC totaled 628 volumes of Chinese material. Among these of the
cataloged works were 454 records for Chinese monographs, which is 72.3% of the items
cataloged in the two month period. Based on the estimation, the analyzed Chinese
monographs (454 records) should occupy at least 20% in the whole academic year of
2006-2007 (approximately 2227 records).

The dates for the creation of the sampled cataloging data in OCLC range from
1980 to 2006. The recent trends in errors can be observed since the major proportion of
the records in this study was created in 2005 and 2006. A total of 16 MARC fields were
examined in this study, they are: 100, 245, 246, 250, 260, 300, 440, 490, 500, 600, 610,
650, 651, 700, 710 and 830.
Chapter IV Data Analysis

A total of 454 records were examined and the error results appear in Table 1:

Table 1: Errors in the Sampled Records

<table>
<thead>
<tr>
<th>Chinese Monographic Records</th>
<th>No. of Records</th>
<th>% of Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>With errors</td>
<td>203</td>
<td>45%</td>
</tr>
<tr>
<td>No errors</td>
<td>251</td>
<td>55%</td>
</tr>
<tr>
<td>Total</td>
<td>454</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 1: Errors in the Sampled Records

The records with errors comprise 45% of the total. The mean error rate for each record was 1.35. For the 203 records with errors, a total of 275 errors were detected. Among the total errors, only 3 errors were recorded in the Error 2 category. This 1% low rate indicated that it is very unlikely to have more than one error in a MARC field.
The detailed description about the errors will be explained from three different perspectives: by content, by type and by MARC field.

4.1 By content

The detailed error descriptions were entered in both Error 1 and Error 2 for further identification and categorization. For the 203 OCLC bibliographic records, a total of 275 errors were recorded in 16 bibliographic fields. Table 2 reflects the main errors with at least a frequency of 4. Other errors with frequencies from 1 to 3 constitute only 14% of the total errors identified.

Table 2: Error Content Analysis

<table>
<thead>
<tr>
<th>Error Content</th>
<th>Fields</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punctuation-Extra period</td>
<td>246,440,490</td>
<td>117</td>
</tr>
<tr>
<td>Incorrect book measurement</td>
<td>300</td>
<td>39</td>
</tr>
<tr>
<td>Fail to provide access point</td>
<td>246, 440</td>
<td>22</td>
</tr>
<tr>
<td>No need for translation</td>
<td>100, 610, 650, 651, 710, 830</td>
<td>12</td>
</tr>
<tr>
<td>Incorrect Pinyin polyphone</td>
<td>245, 246, 260, 440</td>
<td>11</td>
</tr>
<tr>
<td>Incorrect Pinyin spelling</td>
<td>245, 260, 440</td>
<td>8</td>
</tr>
<tr>
<td>Incorrect Chinese script</td>
<td>245, 260, 440, 500</td>
<td>8</td>
</tr>
<tr>
<td>Extra square brackets []</td>
<td>245, 260</td>
<td>7</td>
</tr>
<tr>
<td>Extra spacing</td>
<td>245, 440, 700</td>
<td>4</td>
</tr>
<tr>
<td>Fail to add space separation</td>
<td>245, 246, 300</td>
<td>4</td>
</tr>
<tr>
<td>No parallel Chinese entry</td>
<td>246, 500, 700</td>
<td>4</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>275</td>
</tr>
</tbody>
</table>
The extra period punctuation at the end of the 246, 440 and 490 was the most frequently-seen error in the study. This punctuation existed in both paralleled Chinese vernacular entry and romanization entry. According to Library of Congress’s *MARC 21 Bibliographic Format*, fields 246, 440 and 490 should not end with a mark of punctuation unless the field ends with an abbreviation, an initial/letter, or other data that ends with a mark of punctuation.

An interesting finding in this study is that when fields 246 and 440 appear together in a bibliographic record, if field 246 has the error on the extra period punctuation; it is quite probable for the following 440 field to carry this error as well. In Table 3 the analysis of Field 490 was omitted since only one 490 field was identified with the extra punctuation error and no 246 or 440 field appeared together with 490 field. Thus, the total number of extra punctuation errors in Fields 246 and 440 totals 116.
Half of the errors (58) in Table 3 showing Fields 246 and 440 appeared together. The difference of 2 (30 minus 28) indicated there are two 246 field with the extra period punctuation and the two 440 fields appearing in the same bibliographic records had no extra period punctuation. Thus the extra punctuation error appearing in the 246 and 440 fields together was 97%.

The 300 field (Physical Description) has, in almost all situations, the height of a monograph measured in centimeter (cm.). This is recorded in subfield c (‡c ). Rule 2.5D1 in ACCR2 specifies that the height is recorded to the next whole centimeter. For example, when the height of a book measures 24.2 cm, 25 cm. is the correct height for
recording. The measurement of a book matters when libraries make shelving decisions. In UNC-CH Davis library, books with a height over 30cm. are assigned to the folio shelves.

Fields 246 and 440 are the major fields for providing access points in addition to the title statement entry in Field 245. Both of the former fields contribute to further identification of an item. It is important for catalogers to recognize the varying forms of title or series statement associated with an item. In this study, it was found that it’s easy to overlook the English title, title written in Pinyin, series title at the spine or a corner of a title page. Sometimes, searching for the English title or Pinyin is the only way for catalog users to find a book when the Chinese character input system is not available.

In the sampled bibliographic records, some unnecessary translations were made to the region and country name, e.g., Tibet, China, etc, in fields 610, 650, 651, 710 and 830. Due to the extra translation, a parallel entry in the Chinese vernacular was created.

For example:

650 西藏 (中国) ‡x History.

650 Tibet (China) ‡x History.

should be corrected to the following entry only:

650 Tibet (China) ‡x History.

Another example of this type of error was found in an extra translation in Field 100. Here the Chinese translation should be deleted since this did not follow the authority record:

100 1[陈季同, ‡d 1851-1907.

100 1[ Tcheng, Ki-tong, ‡d 1851-1907.
These entries should be corrected to the following entry only:

100 16 Tcheng, Ki-tong, ‡d 1851-1907.

There are 11 errors found for the Pinyin polyphone. Seven of them are in the field 245. Table 4 reflects an analysis of the Chinese character with polyphone in the bibliographic record:

Table 4: Analysis on Pinyin Polyphone

<table>
<thead>
<tr>
<th>Chinese Character with Pinyin Polyphone</th>
<th>Error Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>的 (di to de)</td>
<td>5</td>
</tr>
<tr>
<td>官 (gong to guan)</td>
<td>2</td>
</tr>
<tr>
<td>乐 (le to yue)</td>
<td>3</td>
</tr>
<tr>
<td>校 (xiao to jiao)</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
</tr>
</tbody>
</table>

Figure 4: Analysis on Pinyin Polyphone
Based on the explanation from http://cjkcataloging.com, the polyphone in the Chinese language refers to a Chinese character pronounced differently depending on the context. Among the 11 errors, the highest rate on Pinyin polyphone errors occurred in the sampled Chinese bibliographic records manifested itself in the pronunciation of the character 的. This character is commonly used in Chinese language.

The Pinyin polyphone error usually appeared in the 245, 246, 260 and 440 fields. In most cases, similar contents on title or part of a title are entered in these fields. Research and practice suggest that it’s possible to have the error copied from one field to another since time spent on repeated typing can be saved when creating a catalog record in OCLC database.

Together with the polyphone errors, errors on Pinyin spelling may cause record retrieval problems. The errors were identified in the 245, 260 and 440 fields. Extra, missing and wrong spelling letters were found. The careless input or lack of proof reading might be the main reasons for the existence of this type of error. For example:

260 郑州 : ‡b 河南人民出版社, ‡c 2004.

should be corrected to:

260 Zhengzhou : ‡b Henan ren ming chu ban she, ‡c 2004.

Eight incorrect Chinese scripts existed in the field 245, 260, 440 and 500. The errors appeared as extra or wrong Chinese characters. Special attention should be paid to the errors since they might affect record retrieval.
When examining the sampled records with square bracket errors, it was found that the content in the brackets can be explicitly identified from the prescribed source of information of the book. The brackets were removed because the symbol “[]” indicated that the content inside the brackets can’t be identified from the main source or it comes from a non-prescribed source of information.

The error of extra spacing and incorrect separation results in the problem of improper separation or connection among Chinese names of authors and regions. The main reason for the error is the failure to follow proper formatting requirements. The errors are likely to appear in both paralleled MARC field entries.

In addition to the *AACR Workbook for East Asian Publications*, there are two important documents to consult to ensure the consistency when creating catalog records for Chinese materials. One is *Romanization of Chinese Geographic Names in Descriptive and Subjective Headings*; the other is *Chinese Romanization Policies for Cataloging Chinese Material*. Both are available on the Library of Congress’s website.

An example of the spacing error on format is as follows:

245 1 0 一九三〇年代鄉土文學: ‡b 臺灣話文論爭及其餘波 / ‡c 陳淑容著.

245 1 0 Yi jiu san ling nian dai xiang tu wen xue : ‡b Taiwan hua wen lun zheng ji qi yu bo / ‡c Chen Shurong zhu.

The error was in the vernacular entry. There should have a space before the colon like the correct format in the romanized entry.

In the sampled records, four of them have no parallel Chinese entry provided in field 246, 500 and 700. Special attention should be paid to those descriptive areas since
the Chinese script data entry in the parallel field can provide more access points for retrieval.

There were 39 other errors are not represented in this study, but rather were found by the author in the course of daily copy cataloging. These include failure to provide edition information, colored maps information, information for accompanying material, name of part/section of a work, alternative title; incorrect date of publication, edition information, pagination, punctuation, and coding of indicators, subfields, etc. Although they are not explained in detail in this paper, attention should be paid to these types of errors since they may cause inconvenience for users.

4.2 By type

Based on the analysis of error content, the errors were further grouped into 5 major types as described in Table 5. No strict borderline was established for each error type since errors might be included in one type or another. The sample was reexamined to determine which type might reflect the main elements of the errors. The 39 errors with frequencies less than 3 were also counted in this type analysis. It is felt that the small portion is unlikely to change the result of the study to any noticeable extent. The coding error types (Table 5) are composed of errors in field code, indicator code and subfield code.
Table 5: Analysis by Type

<table>
<thead>
<tr>
<th>EC</th>
<th>EC 1</th>
<th>EC 2</th>
<th>EC 3</th>
<th>EC 4</th>
<th>EC 5</th>
<th># of Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incorrect description</td>
<td>Incorrect description</td>
<td>Incorrect description</td>
<td>Incorrect description</td>
<td>Incorrect description</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>Book measurement (39)</td>
<td>Pinyin (19)</td>
<td>Chinese script (8)</td>
<td>Square brackets (7)</td>
<td>Other (21)</td>
<td></td>
</tr>
<tr>
<td>Access point</td>
<td>Fail to provide access point (22)</td>
<td>No parallel Chinese entry (4)</td>
<td>Other (5)</td>
<td></td>
<td></td>
<td>31</td>
</tr>
<tr>
<td>Spacing</td>
<td>Fail to add space separation (4)</td>
<td>Extra spacing (4)</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Extra information</td>
<td>No need for translation (12)</td>
<td>Other (3)</td>
<td></td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Formatting</td>
<td>Extra period punctuation (117)</td>
<td>Other (5)</td>
<td></td>
<td></td>
<td></td>
<td>122</td>
</tr>
<tr>
<td>Coding</td>
<td>Other (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>275</td>
</tr>
</tbody>
</table>

EC= Error Content

Figure 5a: Analysis by Type
To get a clear overview of the error types, the 117 errors of extra period punctuation at the end of the 246 and 440 fields were removed since they can be corrected easily. Figure 5b is an adjusted analysis by type. There are only 5 errors on formatting:

**Figure 5b: Adjusted Analysis by Type**

![Pie chart showing error types](image_url)

As can be seen from the above figure, the error types of incorrect description and access points together present about 80% of the 158 sampled items. These kinds of error types play important roles in providing for efficient catalog record retrieval and eliminating confusion due to insufficient description of an item.

The 9% of type of extra information error mainly refers to the unnecessary translation of the fields 610, 650, 651, 710 and 830. Error types in spacing, coding and formatting comprise about 11% of total errors. They may potentially affect proper online display.
4.3 By MARC Field

The error rate in each MARC field is also worth examination. By identifying the principal MARC fields with errors, catalog records can be carefully created with special attention given to the high error frequency fields. Table 6 provides an analysis of all the 275 errors by MARC field:

Table 6: Analysis by MARC Field

<table>
<thead>
<tr>
<th>Field</th>
<th>No. of Errors</th>
<th>% of Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2</td>
<td>0.7%</td>
</tr>
<tr>
<td>245</td>
<td>29</td>
<td>10.5%</td>
</tr>
<tr>
<td>246</td>
<td>68</td>
<td>24.7%</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>1.1%</td>
</tr>
<tr>
<td>260</td>
<td>16</td>
<td>5.8%</td>
</tr>
<tr>
<td>300</td>
<td>44</td>
<td>16.0%</td>
</tr>
<tr>
<td>440</td>
<td>94</td>
<td>34.2%</td>
</tr>
<tr>
<td>490</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>0.7%</td>
</tr>
<tr>
<td>600</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>610</td>
<td>4</td>
<td>1.5%</td>
</tr>
<tr>
<td>650</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>651</td>
<td>2</td>
<td>0.7%</td>
</tr>
<tr>
<td>700</td>
<td>3</td>
<td>1.1%</td>
</tr>
<tr>
<td>710</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>830</td>
<td>4</td>
<td>1.5%</td>
</tr>
<tr>
<td>Total</td>
<td>275</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Fields 246 and 440 are the fields with the highest error rates, 24.7% and 34.2% of total errors respectively. The main reason for this high error incidence is the extra punctuation period error. It alone comprised 44% of the total errors. In Table 6a, the adjusted analysis used for types was employed again to remove the extra punctuation errors in the two fields. Fields with less than 2% of error were grouped together in the “Other” field.
Table 6a: Adjusted Analysis by MARC Field

<table>
<thead>
<tr>
<th>Field</th>
<th>No. of Errors</th>
<th>% of Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>245</td>
<td>29</td>
<td>18.2%</td>
</tr>
<tr>
<td>246</td>
<td>27</td>
<td>17.0%</td>
</tr>
<tr>
<td>260</td>
<td>16</td>
<td>10.1%</td>
</tr>
<tr>
<td>300</td>
<td>44</td>
<td>27.7%</td>
</tr>
<tr>
<td>440</td>
<td>19</td>
<td>11.9%</td>
</tr>
<tr>
<td>610</td>
<td>4</td>
<td>2.5%</td>
</tr>
<tr>
<td>830</td>
<td>4</td>
<td>2.5%</td>
</tr>
<tr>
<td>Other</td>
<td>16</td>
<td>10.1%</td>
</tr>
<tr>
<td>Total</td>
<td>159</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Figure 6a: Adjusted Analysis by MARC Field

Field 245 is used to record the title and statement of responsibility. This is the main information for users employ to identify an item. Field 300 provides the physical description field of an item. Information about any accompanying material is also recorded in this field. For 246 and 440 fields, the problem of access points dominates. Other errors like the incorrect Pinyin or Chinese script may also affect the access to the record.
4.4 Summary of the error findings

Compared to the previous similar study of OCLC records, the quality for cataloging record creation and maintenance is still an issue for Chinese materials. Nearly half of the records in the sample were identified as having errors. However, a large percentage of the errors (42%) can be greatly reduced if the extra punctuation errors in 246 and 440 fields removed from the analysis. This kind of error is not serious since it does not have any negative influence on record retrieval and proper display.

Incorrect description and issues on access point are the errors catalogers least want to see. In this study, aside from formatting errors, these aforementioned errors are the second most dominant error types. They are considered to be serious for their negative influence on catalog records quality.

The major problems are still in the descriptive areas like 245, 246, 260, 300, etc. Since the record for Chinese monographs need double entry effort, the high error rates in these fields are seemingly inevitable. It can be time-consuming to have errors in the fields identified and corrected.
Chapter V Conclusion

5.1 Limitation of the study

There is one major limitation of this study: the lack of ability to conduct a thorough and comprehensive examination of the errors in Chinese monographic records in OCLC. But this limitation can be solved given more time, data, and researchers committed to the issue.

Examination from other aspects of the records such as the fixed field, record entered date, record replaced date, the type of institutions or organizations responsible for record creation and replacement, location of the item publisher, time of publication and other relevant information may help advance the study to determine other possible factors for the existence of the errors, their solutions and methods for correcting the errors.

5.2 Suggestions for future study

Besides the statement about study limitation above for a more thorough study, it should be noted that catalog records for monographs are the primary type of catalog record. Future studies could examine other formats for errors since there are different requirements for record creation and they tend to be much more complicated.

5.3 Possible measures for eliminating the errors

The purpose of this paper was to examine some characteristics of errors in the OCLC catalog records for Chinese monographs. During the correction process for the
sampled records, several possible measures for eliminating the errors had been identified and could be implemented at record creation, copy cataloging or later maintenance phases. Several recommended measures are already in use which could, if enhanced, provide the capability to more immediately correct and notification of errors, thus reducing the cost of unnecessarily repetitive labor costs.

1. Efficient methods to notify the cataloging changes

   In addition to the existing cataloging rules and formatting documentation, amendments and changes are often made for cataloging Chinese materials. Efficient methods are needed to notify catalogers of essential changes to ensuring the consistency of the catalog record and implementing changes.

2. Development of an effective procedure for catalog record checking before final uploading to the OCLC WorldCat

   Aids from both people and the computer are needed. It is quite likely that errors can be identified if a second person does rechecking after the creation of a record by a first person. Automatic validation on MARC tags and codes has been adopted by OCLC as a tool to decreasing the errors in MARC formatting.

3. Extensive training on the cataloging of Chinese materials

   The training will be a dynamic activity for Chinese material catalogers. New changes, trends and related materials which may facilitate Chinese materials’ cataloging can be shared. Concerns and ideas on better cataloging practices should be encouraged since many Chinese cataloging documents are still under revision. All persons involved in cataloging and copy cataloging should
participate in this training, especially student assistants. Ensuring their deep understanding and high proficiency will help to improve the quality and reliability of the catalog records.

4. Referencing tools

Many errors on Chinese character pronunciation and the writing of Chinese scripts can be greatly reduced provided there is a conveniently accessible referencing tool to consult. In most cases, a bulky Chinese dictionary can be put away if an authoritative online Chinese dictionary is available and it can be accepted in the library community at large.

5. Extra emphasis should be placed on the errors that affect record access

An online catalog made detailed descriptions of items possible, but different descriptive levels of an item not only affect access to that specific item, but also of access to the group of items under its specific categories. The accessibility of the online catalog records is one of the most major concerns for libraries. Extra attention needs to be given towards selecting access points which anticipate users’ search habits.

6. Enhanced error reporting system

OCLC has implemented error reporting systems for its member libraries. Detailed instructions are available at:

http://www.oclc.org/bibformats/en/quality. However, there are requirements for authorization to change a record. Some report submitting methods may take time for OCLC to process. Less serious errors are often corrected by local libraries for their local display. Thus, they may not want to spend extra time to
file an error report. A potential ameliorant to this problem is making available an enhanced error reporting system that allows for more immediate reporting and correction by providing a list of shared errors and of how to fix them, which is reviewed and managed by OCLC, for catalogers’ reference.
References


Doszkocs, T. E., et. al. (2004). Dictionary Services and Spelling Aids for Web Searching. *Online (Weston, Conn.)*, v. 28 no. 3, p. 16-21


WorldCat database CJK statistics, Accessed Feb 12, 2007,  
http://www.oclc.org/worldcat/statistics/cjk.htm