Amigos Fellowship Final Report for

A Quick Cataloging Guide for CJK Catalogers: Analysis of the Most Frequent Errors in CJK Bibliographic Records

Kai Yu

Texas A&M University Libraries
College Station, TX 77840

December 2006

Final Report to the Amigos Fellowship Program-2005

Report for A Quick Cataloging Guide for CJK Catalogers: Analysis of the Most Frequent Errors in CJK Bibliographic Records

This final report is respectfully submitted to the Amigos Fellowship Program, Amigos Library Services as a competition of abovementioned project. This final report consists of four parts: Background, Summary of Project Activities, Financial Statement, and Evaluation and Conclusion.

Background

The world has seen a fast expansion of CJK bibliographic records in recent years. According to the CJK statistics reports provided by OCLC, the number of CJK language-coded records has grown from 1932,412 in January 2001 to3,189,123 as of January 2006. The growth rate is 39.4% over the five-year period.

The rapid expansion of CJK resources reflects the booming interests in substantial business and other opportunities in East Asia. However, the relatively high error rate in CJK bibliographic records has been a big hurdle in helping patrons to access this unique source of information resources, and has reduced efficiency in library cooperation and resource sharing. This problem has drawn much attention from library professionals, particularly from catalogers.

This project aims to provide a thorough and updated analysis of the most common errors in CJK bibliographic records drawn from the OCLC database. In this research, 1,000 sample records were randomly selected; the errors were identified and categorized; and analyses were performed to explore why the errors occurred and how to improve the quality of the CJK records.

Summary of Project Activities

The implementation of project is decomposed into the following five steps:

Step 1: Data Collection

This is the most time-consuming, but very critical activity for the project. Up to 600 hours have been spent in this activity. A sample of 1,000 CJK bibliographic records has been randomly selected from the local Voyager Database. All the records were imported directly from OCLC WorldCat. The following selection criteria were employed:

a) Choose the records with language coded "chi" (Chinese), "kor" (Korean) or "jpn" (Japanese) in the fixed field. This procedure ensures that the selected publications are in Chinese, Korean and Japanese.
b) Choose the records containing field 880 to ensure both the vernacular fields and the Romanized fields are included in the records.
c) Select records that include field 987 with letter "c" in subfield $d for Chinese bibliographical records. This step is to ensure that the sampled records have been fully romanized, contain pinyin romanization for Chinese scripts, and thereby eliminate the erroneous conversion.
d) All sampled records should have the following bibliographic characters: "a" for Type, "Print" for Format, and "I" for Bibliographic Level. For research purpose, all the selected records are in print book format.

Step 2: Data Analysis and Error Classification:

Up to 300 hours have been spent in this activity. Sample records collected in Step 1 have been thoroughly examined against the physical pieces in hand. Errors are identified, and details are saved in an Access file. According to the frequency of error occurrence, ten types of the most common errors have been categorized. Listed below is the summary of the identified errors and the number and rate of occurrence:

Figure 1: Summary of Errors in CJK Records

Error Category	Number of Occurrence	Error Rate
Incorrect Separation or Connection of Syllables	97	9.7%
Mixed Data Problems	71	7.1%
Incorrect Punctuation	12	1.2
Incorrect Capitalization	18	1.8%
Incorrect Pronunciation	45	4.5%
Typos and Mismatches	24	2.4%
Incorrect Diacritics and Special Character	13	1.3%
Absence or Inaccuracy of Tag Fields	18	1.8%
Errors in Punctuation (MARC)	52	5.2
Errors in Parallel Fields	20	2.0%

Note: An active linkage has been established for each of the above listed error categories. By clicking each error category, you will find all detailed analyses and example records from the web-based aid.

Step 3: Code Writing and Domain Registration:

Up to 268 hours have been used in code writing, web design, and selecting an appropriate service provider to host the web-based product of this project - the *Interactive CJK Cataloging Aid*. Please check the site for more details (http://cjkcataloging.com).

Step 4: Proof-Testing Activity:

Up to 200 hours have been spent in proof-testing. I chose Texas A&M University free web space for faculty as my "test field", and uploaded the web-based project there (http://people.tamu.edu/~karenyu) for peer catalogers' review and comments.

Step 5: Post-Project Activity:

This is the cleanup stage to finalize the project outputs. Up to 100 hours have been used to make corrections based on the feedbacks received, and the mistakes identified in earlier stages.

Financial Statement

The certified Financial Statement is provided in a separate PDF file.

Evaluation and Conclusion

According to the proposal, the draft of the *A Quick Cataloging Guide for CJK Catalogers: Analysis of the Most Frequent Errors in CJK Bibliographic Records* was first uploaded to TAMU's web space (http://people.tamu.edu/~karenyu) for review and comments by CJK catalogers and experts before it was published to the public web space (http://cjkcataloging.com). The comments and feedbacks have been quite positive to this project. In addition to a thorough and updated analysis of the most common errors in CJK bibliographic records, I expanded the scope of the original proposal by adding five more components to this project. The additions make it "very practical and comprehensive" according to the comments by a colleague in the New College of Florida State University. A new name "*the Interactive CJK Cataloging Aid*" has been provided accordingly to reflect the expanded scope of the project.

The project outputs will benefit the CJK cataloging community. It not only advances my cataloging skills, but also provides more opportunities to develop my career and research. After reviewing my project, one CJK expert at Columbia University recommended that I make a formal presentation at CEAL annual conference in the coming year. Another CJK expert at Duke University in Durham was very impressed upon my research, and we have started teaming up for a new research project.

Amigos Fellowship program provides me with generous financial support that enables me to make my dream research come true. I gratefully acknowledge and appreciate their support.