The Pinyin Conversion Project and the Challenge of Cleaning Up Afterward

Philip Melzer

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

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/jeal/vol2022/iss175/5

This Report is brought to you for free and open access by the 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 ellen_amatangelo@byu.edu.
The Pinyin Conversion Project and the Challenge of Cleaning up Afterward

Philip Melzer
Library of Congress (ret.)

It has been more than 20 years since the beginning of the Pinyin Conversion Project. At the time, this massive cooperative effort was the largest library data conversion following the adoption of AACR2 in 1981 and the conversion of LC’s entire set of databases to load into the new ILS in 1999. I served as coordinator of the project at the Library of Congress (LC). This paper gives an outline of the project, from my point of view at LC. These are my own views and not a representation of LC policy.

Overview of the American conversion

Here is a brief overview of the project.

- In 1996, the National Library of Australia used a machine program to convert its roughly 500,000 Chinese bib records.
- In November 1997, LC announced that it intended to convert to pinyin Romanization within the next several years. The bibliographic utilities OCLC and RLG agreed to conduct the conversion itself.
- Moratoriums on Chinese language authority work and the cataloging of Chinese material were called in August 2000, to prevent work on records undergoing conversion.
- In September 2000, just prior to the official changeover to pinyin, OCLC converted 158,368 name authority records. By November 2000, LC had completed the conversion of roughly 500 subject authority records that included romanized Chinese.
- Day 1 for pinyin occurred on October 1, 2000. After that date, pinyin became the standard for Chinese Romanization in American libraries.
- OCLC completed the conversion of approximately 700,000 Chinese master records in June 2001, including roughly 9000 LC serial records.
- RLG began converting Chinese bib records in November 2000 and completed the conversion of over 2,500,000 records by December 2001, including 172,487 LC Chinese language bib records.
- Between 2000 and 2005, LC staff manually reviewed or converted several thousand name authorities and approximately 45,000-50,000 bib records.

What romanization is

One of the first, and most vital steps in the project, was to come to a common understanding of what romanization is, and what it is not.
Romanization is the representation of non-roman languages in roman letters, which seeks to represent the sound of the language, as spoken, or how the language is written, or a combination of the two. It is a representation of a non-roman script, and not the language itself. Even though we have included certain non-roman scripts in bibliographic records for many years, we still rely on Romanization to make it possible to file and retrieve non-roman data in our roman-letter-based bibliographic and authority files.

Romanization schemes used by libraries are developed by LC, or by other institutions or individuals who have linguistic expertise that LC lacks. Tables are prepared, discussed, and evaluated by LC, the Committee on African and Asian Material (CC:AAM) of the American Library Association (ALA), or language and subject experts. Finally, they are then approved by LC and CC:AAM. The compendium of Romanization tables is the ALA-LC Romanization tables, available on the LC website: https://www.loc.gov/catdir/cpso/roman.html

Here is an example of an RLIN Chinese bib record, updated in 2004. Most of the romanized fields are followed by parallel fields with corresponding non-roman data.

Pinyin romanization

Between the 1950s and October 1, 2000, the standard for Chinese Romanization in the U.S. and other countries was a system called Wade-Giles. A different system called pinyin became the standard Romanization scheme in China in the late 1950s, and gradually became the scheme that was used and accepted internationally. LC wanted to convert to pinyin at the same time it began creating East Asian cataloging records online, in 1980. However, the East
Asian library community did not support the conversion at that time, as reflected in a vote on the matter at the Committee on East Asian Libraries (CEAL, later the Council on East Asian Libraries). As a result, LC did not convert or move to adopt pinyin. By 1984, pinyin had been adopted by the Federal government as the standard for romanizing Chinese, but the East Asian library community still did not show its support. In 1990, LC decided that it would have to convert to pinyin when technology made it possible to do so comprehensively and accurately. Several more years were to pass before this became a reality.

Pinyin has several advantages over Wade-Giles. Because of its worldwide acceptance and use, library patrons were familiar with it, and it facilitated the international exchange of data.

**Pinyin makes clearer distinctions between syllables than Wade-Giles**

<table>
<thead>
<tr>
<th>character</th>
<th>Wade-Giles</th>
<th>Pinyin</th>
</tr>
</thead>
<tbody>
<tr>
<td>図</td>
<td>t‘u</td>
<td>tu</td>
</tr>
<tr>
<td>读</td>
<td>tu</td>
<td>du</td>
</tr>
<tr>
<td>出</td>
<td>ch‘u</td>
<td>chu</td>
</tr>
<tr>
<td>朱</td>
<td>chu</td>
<td>zhu</td>
</tr>
<tr>
<td>去</td>
<td>ch‘ü</td>
<td>qu</td>
</tr>
<tr>
<td>句</td>
<td>chū</td>
<td>ju</td>
</tr>
</tbody>
</table>

Wade-Giles frequently used the same letters to represent different sounds, while pinyin makes more distinctions and presents few ambiguities by using all the letters of the roman alphabet. Wade-Giles used diacritic marks frequently; pinyin uses only the umlaut, on 4 syllables—making pinyin easier to input, easier to read, and less likely to involve typographical error.

**The Australian conversion**

In 1995, we at LC learned that the National Library of Australia (NLA) intended to convert its bibliographic records to pinyin. We got in touch and were invited to review the test file before their conversion, and then converted records afterward. In 1996, NLA began to use a computer program to convert, in the end, over 500,000 bib records, including all the Chinese, Japanese and Korean (CJK) records made available by the LC’s Cataloging Distribution Service (CDS). Their program analyzed fields by MARC tags. About 12% of the converted records were set aside for manual review. The high quality of their conversion demonstrated that machine conversion was possible and gave us hope that we too might be able to convert our records.
Early planning

Encouraged by the accomplishments in Australia, in 1997 LC staff began meeting with staff from the two major bibliographic utilities, OCLC and RLG, to discuss how we could cooperate and convert to pinyin. We agreed that such a task would be risky and complicated: the conversion would affect libraries throughout the country; and would involve changing hundreds of thousands of records in the same manner in our very interconnected files.

Nevertheless, in November 1997, LC announced that it believed it was now feasible to convert to pinyin and would do so in the following several years. RLG, and then OCLC, agreed to convert the romanized portions of the Chinese records in their files. OCLC generously agreed to convert name authority records. The agreement on the part of the utilities was absolutely necessary for this national effort and had the effect of bringing other institutions along.

A cooperative effort between LC, the major bibliographic utilities, major libraries, and library organizations

Gradually there was recognition that the conversion project was going to affect everyone. Even small collections and public libraries had romanized Chinese in their catalogs - headings for Chinese people, place names, and subjects. Through public forums and programs at professional meetings, LC and the utilities attempted to reach beyond the East Asian community to prepare everyone for the conversion and to seek buy-in from the larger library community.

By the summer of 1999, the mechanics of the conversion itself were taking shape, but many aspects of the project had not been worked out. How would the project coordinate efforts between institutions? In what order would conversions of authority and bib records occur? How would converted records be identified? Would this lead to split files?

On September 7, 1999, a planning meeting was convened at LC at the behest of Jeffrey Horrell and Dale Flecker of Harvard University. Representatives of six major collections, OCLC, RLG, and LC attended. A consensus was reached on desired outcomes, and on a collective plan of action. Several milestones were agreed upon: Day 1 would constitute the official changeover to pinyin for all library operations at LC. The conversion of authorities would be completed on that day, and the conversion of bib records would begin as soon after as possible. The second milestone would occur when an individual library’s records had been converted; this would be a different date for each library. The third milestone would be declared when it was no longer necessary to add a pinyin marker to records.
A general timeline was worked out. Before Day 1, there would be agreement on markers for authorities and bib records. After Day 1, new authorities would be established in pinyin form. Converted subject authorities would be released as soon after Day 1 as possible. LC's Beacher Wiggins, then LC Director for Cataloging, declared that Day 1 would take place on October 1, 2000, providing a target for everyone to work toward. Issues related to cleanup were set aside. The utilities agreed to convert the Chinese records in their files, to convert other files that were sent in by member libraries, and return snapshots of converted records to the library of origin. Standards for accuracy, cost of services, and cataloging moratoriums were discussed. LC and the utilities were urged to publicize the conversion timeline and communicate as much about the project with the library community as possible.

There was, from my perspective, harmonious cooperation between LC and the utilities, and generous sharing of information concerning pinyin conversion between the parties. RLG, OCLC, and LC managers and staff held conference calls at least monthly for approximately 2½ years, to agree on courses of action, coordinate activities, and plan the conversion timeline. The project was also supported and aided by well-informed and involved task groups that were formed at LC, OCLC and RLG, and at CEAL. In each of these institutions, smaller teams were formed to plan and accomplish the conversion.
LC established a Pinyin Home Page which was updated frequently with accomplishments and timeline changes. The page included as many documents and as much information about the pinyin conversion project as it was possible to supply, including romanization policies and explanations; FAQs; conversion specifications; and information about markers, subject headings, and classification. Even today the Pinyin Home Page can be viewed in its entirety at: www.loc.gov/catdir/pinyin.

Pinyin romanization guidelines

Pinyin romanization guidelines were drafted in 1998, following Chinese guidelines on most points. The library community was asked for input. The most important - and most emotional - decision was whether or not to aggregate syllables and form lexical units. Because Wade-Giles Romanization separated syllables, it would have been far too complicated to try to determine which syllables should be connected to represent lexical units, especially since records covering all subjects and time periods would be converted, and tens of thousands of records did not include parallel Chinese text. In addition, many institutions had records in their files that employed pre-Wade-Giles Chinese romanization practices. Therefore, although the Chinese guidelines called for the aggregation of syllables, it was decided that the practice of separating syllables would have to continue in the LC-ALA guidelines.

Interestingly, LC staff met with representatives of the National Library of China in the summer of 1998 to discuss NLC’s Chinese romanization and word division practices. That library also used romanization for purposes of filing and retrieval, just like we do, and their practice was to separate all syllables.

After the adoption of pinyin romanization guidelines, and for several years following the conversion, LC took advantage of the opportunity to address all Chinese romanization issues.
and undertake the extensive task of promulgating detailed written procedures and correcting various inconsistencies that had taken hold over the years.

Conversion specifications

Authority records give the authorized form of name for use in headings, cross-references leading to the heading, and sources for this form and variant forms. Headings in the form found on authority records are used on bib records to facilitate searching and the collocation of personal and corporate names, names of meetings, geographic locations, titles, and subjects. They are one of the basic elements of bibliographic control.

Authority data in the National Authority File was shared between LC and the utilities, OCLC and RLG, and the British Library. The data flowed back and forth and was kept up to date on a daily basis. A great deal of effort was expended at each institution to always keep the files in sync. Because authority control was the basis of the interrelationship between authority and bib records, the conversion project had to be coordinated so that the results of the conversions would be the same. An authority record could not be allowed to convert one way, and a corresponding heading on a bib record another.

Conversion specifications (specs) were drafted by a group at LC, working closely with staff at OCLC and RLG, and with an RLG working group made up of librarians with expertise in Chinese materials. The specs had to specify exactly what character strings were to be converted by machine for both authorities and bib records. They were based on assumptions of what one was likely to find in each subfield of each record that was a candidate for conversion. The specs identified and analyzed systematically romanized Chinese syllables so that the computer program knew when and how it should convert, when it should not convert, and when it had to mark a record for review. The specs defined and described in exhaustive detail all exceptional situations (personal names, geographic names, dates, generic terms) and then the fields and subfields in which they would be found. They determined the exact sequence of conversion of each subfield. On top of that, they had to be written in such a way that programmers at RLG and OCLC could use them to create several very different conversion programs that would achieve identical results: one against the RLG Union Catalog, one against the LC/NACO Name Authority File, and a third against the OCLC WorldCat database.

Specifications relied primarily on a combination of data dictionaries and conversion sequences. Data dictionaries included:

- standard conversion
- syllables - definition (WG, PY, same, common, other)
- conventional place names (3)
- personal name exclusion list
- multi-syllable generic terms for jurisdictions
- chronological subdivisions (bib records only)
- subject headings (corporate, geographics (2), topical) (in bib records only)
Conversion sequences included:

- mixed text
- personal names
- single syllable generic terms for jurisdictions
- Taiwan place names (3)
- 2-syllable place names
- references (in converted authorities only)

**Specifications for conversion of name authority records**

The conversion of name authorities was difficult because authorities were not then, and are not now, identified by language. The scanning program had to consider the entire name authority file. Candidates for conversion included headings with one subfield identified as Wade-Giles. Pre-AACR2 headings were not converted. Capitalization and punctuation generally did not change. On authority records that were converted, references for personal names and place names were converted, and the Wade-Giles forms were retained; however, references for corporate bodies, meetings, and uniform titles were converted, but the Wade-Giles forms were not retained. The 13% of the headings that were identified by the program as being ambiguous - that is, they could either be Wade-Giles or pinyin - were assembled in files and sent to LC for review. An example of the procedures for converting just the $a subfield of field 130 is shown here:
The specifications described Wade-Giles romanization in great detail, particularly personal names and geographic names. The basic procedures for conversion were outlined: first, the scanning procedure by which candidates for conversion were identified; then, the order of conversion, field by field. Personal names were run against an exclusion list. Then each subfield was analyzed, and data dictionaries and conversion sequences were run in a certain order for each subfield. At each step in the process, the program was directed to pass, convert, or convert and send to review file, and then de-dup headings.

The following example shows the converted name authority record for Xie Lingyun. The heading was converted, and the former heading (in Wade-Giles form) was retained as a reference. The two see references in Wade-Giles form were converted to pinyin, and the Wade-Giles forms were retained. Non-Wade-Giles references were unchanged.
Name authorities were converted before bib records because authority control is basic to so many library systems and we wanted these authority records to be available for use on Day 1.

Markers

There was general agreement that it would be necessary to mark a record when it was converted, in order to quickly and conveniently distinguish it from a record that had not yet been converted. This was a vital safety feature. The utilities were processing a huge volume of records by machine and wanted to make sure that they would not convert the same record twice. Marked records would also facilitate cleanup tasks.

After wide consultation with the library community and vendors, it was agreed that the 008/07 field would be used by the conversion program to mark whether an authority record had been converted or considered for conversion and passed over. The 987 field was used to mark bib records with information identifying the converting institution, date of conversion, and what if anything in the record needed to be reviewed. 987 fields were used until the year 2008 when the utilities agreed they were no longer needed.

Personal names

Librarians were particularly concerned that a machine program would convert headings for personal names that might appear to be romanized, but were in fact transcribed – and should therefore be left alone. OCLC and LC staff worked together to reduce the likelihood that
personal name headings would be mistakenly converted. After perfecting the identification of personal names in Wade-Giles form, OCLC staff scanned personal name headings on LC records in WorldCat. They analyzed data on both authority and bib records on which the headings appeared. Then, following several menus, they created eight lists of the headings in Wade-Giles form most likely to have been transcribed and not romanized. That information was analyzed to determine whether it was appropriate to convert the heading or not. As I recall, roughly 20,000 authority records were sent to LC and analyzed by hand by Beatrice Ohta and myself. Headings that were not to be converted were then put into a file called the Exclusion List, which ended up including more than 2400 headings.

After collecting candidates for conversion, the first step in the conversion of both authorities and bib records was to run personal name headings against the Exclusion List. If a heading on an authority record matched one on the list, the 008/07 was marked ‘n’ for ‘do not convert’ and the heading was passed. If a heading on a bib record matched one on the exclusion list, it was passed and not converted.

The Exclusion List proved to be very effective. In the 4½ years following the conversion, LC staff encountered at most a few dozen names that had been erroneously converted.

**Conventional place names**

More than 150 headings for Chinese place names, known as “conventional place names” because they were not romanized according to the Wade-Giles scheme, were folded into the project. Hundreds of authority records involving these headings were converted manually at LC, beginning in 1998 in anticipation of pinyin conversion. A few of these headings on authorities were left to be converted by machine. All headings on Chinese bib records were converted by machine. Three data dictionaries were used to convert the names appearing in different forms in different parts of a heading. A chart listing all of the converted conventional names appears on the Pinyin Home Page.
<table>
<thead>
<tr>
<th>Former conventional form</th>
<th>Converted form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peking (China)</td>
<td>Beijing (China)</td>
</tr>
<tr>
<td>Port Arthur (China)</td>
<td>Lushun (China)</td>
</tr>
<tr>
<td>Canton (China)</td>
<td>Guangzhou Shi (China)</td>
</tr>
<tr>
<td>Hupeh Province (China)</td>
<td>Hubei Sheng (China)</td>
</tr>
<tr>
<td>Sinkiang Uighur Autonomous Region (China)</td>
<td>Xinjiang Uygur Zizhi (China)</td>
</tr>
</tbody>
</table>

Taiwan place names

Because at the time of conversion Taiwan continued to use Wade-Giles Romanization, the US Board on Geographic Names (BGN) could not change the official form of place names in Taiwan to pinyin form. Then, in turn, because the Library of Congress was obliged to follow BGN's official forms for naming places, the conversion specs had to describe how to distinguish headings for Taiwan place names from other place names on authority and bib records (in headings, in qualifiers for localities and meetings, in subject subdivisions). In each subfield that could include such a heading, a scan was called for and the conversion programs blocked the Taiwan place name from conversion. This portion of the conversion programs was difficult to put into practice but proved to be highly successful in blocking from conversion the headings that included Taiwan place names.

Here is an RLIN bib record in which all Wade-Giles romanized text converted except for the heading for the Taiwanese city.
Specifications for conversion of bib records

Bibliographic records posed a different challenge than authority records, in that they tended to include many more fields and subfields. All records coded Chinese in the RLG Union Catalog and OCLC Worldcat databases were collected and converted. In addition, libraries sent files of Chinese records to the utilities for them to convert. (Some Chinese records were not sent to be converted – for example, LC’s PREMARC Chinese records.)

Conversion specifications included instructions for any subfield that could possibly contain romanized Chinese. Topical subject headings were converted from a list in a data dictionary. Headings for personal names were first run against the exclusion list; if the heading was not excluded, it was converted only if it appeared in full Wade-Giles form. All records were marked with a 987 field.

Mixed text

One of the most vexing problems we dealt with was the fact that, in certain subfields in bib records, romanized Chinese was sometimes mixed in with other non-Chinese syllable strings. How could the machine program be directed to convert as much of it as possible, and then mark the record if some questionable text was left unconverted? How could the conversion of non-Chinese syllables in mixed text best be avoided?
A data dictionary and conversion routines were combined to convert mixed text in certain subfields. First, if possible, the subfield was broken down into smaller units between punctuation marks. Then, syllables in each unit were analyzed against a data dictionary and classified as being 1) unique Wade-Giles, 2) same (Wade-Giles and pinyin use the same letters to represent the same sound), 3) common (Wade-Giles and pinyin use the same letters to represent different sounds), 4) unique pinyin, and 5) other. As you might imagine, the presence of 69 common syllables complicated conversion programming a great deal.
The program then ran the syllables against a mixed text conversion sequence to determine whether that unit should 1) be converted; 2) be left unconverted, while marking the record; or 3) be passed.

<table>
<thead>
<tr>
<th>WHEN TEXT IS</th>
<th>TAKE THIS ACTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>consider WG = convert</td>
</tr>
<tr>
<td>23</td>
<td>consider WG = convert</td>
</tr>
<tr>
<td>13</td>
<td>consider mixed = mark</td>
</tr>
<tr>
<td>24</td>
<td>consider mixed = mark</td>
</tr>
<tr>
<td>34</td>
<td>consider PY = pass</td>
</tr>
<tr>
<td>45</td>
<td>consider other = pass</td>
</tr>
</tbody>
</table>

The program then ran the syllables against a mixed text conversion sequence to determine whether that unit should 1) be converted; 2) be left unconverted, while marking the record; or 3) be passed.

**MIXED TEXT SCANNING ACTION**

1) analyze subfield  
2) break into smaller sections  
3) analyze each section, converting it if possible  
4) analyze entire subfield again  
5) take action:  
   a) if entire subfield has now converted, move on  
   b) if any portion of the subfield is identified as mixture of WG and other, mark and move on  
   c) if portions consist of converted sections and non-WG syllables, move on without marking

After analyzing each unit in the subfield, the program analyzed the entire subfield again. If the whole subfield converted, the program moved on; if portions of the subfield were identified as a mixture of Wade-Giles and others, the record was marked; if portions consisted of converted sections and non-Wade-Giles syllables, the program moved on
without marking the record. By using the mixed text routine, the conversion programs were able to convert thousands of records that would otherwise have been marked for review.

**Testing, testing, testing**

LC compiled test files of approximately 150 authorities and 150 bib records for use by RLG and OCLC. The test records were selected because they would present the conversion programs with ordinary situations as well as exceptions and challenges. OCLC ran 7 tests of authorities, 11 tests of bib records, 3 tests of CONSER records, and 4 tests of non-Chinese bib records. RLG conducted 9 tests of bib records.

A test consisted of running the most recent version of the conversion program against the file of test records. After each test, a file of the converted test records was sent to LC, where results were immediately analyzed by me and Beatrice Ohta. Each letter and word of each record was reviewed. Then, on short turnaround, a detailed report was written—what parts of the program worked, and what parts didn't. Frequently we would then conduct a conference call with RLG or OCLC staff to analyze the results and discuss problems with the specifications and possible program changes. Based on the results, programmers updated their programs and LC adjusted the specs. Then the utility would schedule another test, and the cycle would be repeated. A great deal of daily—indeed hourly—correspondence took place during the testing period. It took a long time to lock down the program and specifications for each conversion.

**Subject headings and classification**

An LC subject cataloger, with help from the Cataloging Policy Office, hunted down subject headings that included terminology in romanized Chinese. He recommended which headings should convert, and which references should be retained. Policy Office staff then made the final decision. That office compiled a list of headings in Wade-Giles and the corresponding converted pinyin form for data dictionaries in the conversion specifications. Shortly before the date of the machine conversion, the Policy Office began manually converting the subject authority records.

Names and terms throughout the printed classification schedule were manually changed to include or convert to pinyin Romanization. Conventional Chinese place names were changed throughout the schedule. Generally, Cutter numbers printed in the class schedule were retained. Then, the rule for the future became: Cutter what you see, even though editions may be separated on the shelf. As you might imagine, this was a major decision with far-reaching consequences.

Several major changes were also made to the class schedules. For example, in class PL, language and literature, the period 1949-2000 was closed for Chinese literary authors and a new chronological period was initiated, beginning with 2001-. The most significant change was made in Class DS, Chinese place names.
Because they followed the Romanization scheme, Cutter numbers for Chinese material under Wade-Giles romanization guidelines tended to accumulate at a few locations in the alphabet. Because LC acquires so much material about Chinese local history, the Cutter numbers for some Chinese localities had reached six and seven numbers. How could this be avoided in the future?

A new class number, DS797, was created for all places smaller than provinces. The number was subdivided artificially by province, rather than sub-classifying strictly alphabetically by the name of the locality. For example, DS797.22 was used for Anhui Province, .24 for Chongqing, .26 for Fujian Province, and so forth. This had the effect of creating space between the numbers for localities from the beginning. The decision to use 3 numbers in Cutters also had the effect of postponing the need for lengthy Cutters.

Consider this example. The book is about Fuzhou Shi, the capital of Fujian Province. DS797 is the class number for Chinese local history; .26 is the sub-class for Fujian Province; .F89 is the sub-classification for Fuzhou Shi; F895 is the Cutter for the title of the book; 2000 is the date of publication.

The establishment of the classification scheme in DS797 proved to be one of the most significant achievements of the conversion because it had the effect of imposing increased orderliness upon an important, crowded, and heavily used portion of the LC collection.
Testing and loading records at LC

Careful and thorough testing was one major way LC managed risk to its operations. The LC database was the primary record of the collections and was used daily by hundreds of LC employees. Therefore, tests were conducted in an environment as close to actual production as possible. The pinyin data loads were a stern test for LC’s new ILS. The conversion, record structure, and data load were first tested in a test database, to simulate all stages of the record transfer process. Records were then loaded into the LC database and reviewed again. Many LC staff had roles in the testing and loading of the test records and the converted records, following complicated testing schedules and meeting stringent deadlines. All converted authority records were successfully loaded over a single long weekend. The loading of converted bib records took many days because of competing priorities, and the huge number and complexity of the bib records.

Cleanup

Pinyin conversion resulted in mixed catalogs, both at the utilities and individual institutions. But it proved difficult to share the burden of many aspects of pinyin cleanup. RLG proved unable to find a way to avoid having many different institutions correct the same record in a cluster. In 2008, OCLC asked the libraries that created marked non-Chinese records to finally call them up and correct them.

On average, between 12% and 15% of each institution’s records were marked for review. More than 21,000 LC bib records were marked, about 13% of the total number. Analysis showed that one-third of the records were marked because of typos in the original record, and another third because of mixed text.

Cleanup at other institutions

Some libraries, such as Michigan and Yale, had integrated systems that enabled updates by machine program after the file of converted pinyin authority records was loaded. But many other institutions were forced to carry out many cleanup activities manually. Princeton, like LC, had tens of thousands of brief roman-only records which were not machine converted because of low quality and non-Wade-Giles romanization schemes.

Initial phase of cleanup at LC

At LC, Chinese language catalogers reviewed and cleaned up converted authority records for several months immediately following their conversion. A special project to clean up some 8400 undifferentiated Chinese name authority headings was assisted by volunteers from eleven NACO libraries; that process took several years to complete.

Bib records that were marked for review were divided by priority—those with possible errors in access points were reviewed and corrected within weeks of the completion of conversion. Because of a shortage of resources, many records with errors in non-access points were either marked “access not affected” or were left alone.

Cleanup of unconverted records

All of the romanized Chinese that could not be found and converted by software applications had to be found and converted manually. There were thousands of unconverted headings on
PREMARC Chinese and non-Chinese records—headings for famous people and place names, uniform titles of translations, subject headings, and chronological subdivisions.

**LC’S PINYIN CLEANUP TASKS**

**Authority records**
- Converted records
  - Files of converted authority records for review from OCLC
- Unconverted records
  - Undifferentiated personal names
  - Uniform titles under personal names, corporate bodies

**Bib records**
- Converted records
  - Marked records
  - Conversion errors - double conversion, multi-syllable terms
  - Personal names – religious titles
  - Unconverted subject headings
  - Subject headings for regions
  - Generic terms for geographic features and jurisdictions
  - Headings using the term Minguo
- Unconverted records
  - Subject headings
  - Most frequently used headings
  - Conventional place names
  - Non-book formats (instrumental music, motion pictures)
  - Generic terms for geographic features and jurisdictions
  - Mongolian and Tibetan records
  - “Title in Chinese”
  - 041 and 043 scans
  - Well-known authors of works in Chinese and English
  - Headings using the term Minguo

Several different cleanup strategies emerged from meetings of LC staff. Database searches were devised to target clusters of records that probably contained some of the headings with unconverted romanized Chinese strings that were likely to be accessed by users. Some cleanup strategies involved targeting records in certain formats or appearing in certain languages. Because, at the time, LC did not then have a mechanism for handling global heading changes in its database, these cleanup tasks were accomplished manually. In most cases, a search was conducted; records were called up manually from the search results, analyzed, and converted, in part or in full.

A list of the most frequently used Chinese headings in WorldCat, provided by OCLC, was used as the starting point for conversion of the 164 most used headings, on a total of 17,000 LC bib records.

**Conclusion**

In my view, considering the complexity of the task and the number of different institutions involved, the Pinyin Conversion Project was highly successful. In the end, the results were highly accurate. The marker on bib records effectively identified and prioritized the records, and the fields within those records, needing change or review. The percentage of authority and bib records that resulted in conversion errors was, in the end, minuscule when
compared to the vast number that converted correctly. It was recognized from the outset that the conversion project would be difficult and would consume time and resources over a period of several years. Nevertheless, procedures and targets were devised and agreed to, and then the conversion was planned and carried out on a broad, national scale. There was effective and far-sighted leadership; widespread cooperation among institutions; and diligence and professionalism shown by administrators, librarians, IT specialists, and support staff, between and within many institutions. I think that the project showed us at our best.