Library of Congress
Pinyin Conversion Project

Conversion and Cleanup Tasks: Status Report
June 4, 2003

COMPLETED TASKS

1. Review of files of converted authority records

While converting authority records, OCLC created a number of files containing certain kinds of records for review. For more than five months in late 2000 and early 2001, LC catalogers painstakingly reviewed the thousands of converted authority records in these files and made corrections where warranted.

2. Authority records for undifferentiated personal names

The conversion program did not convert authority records that were coded as undifferentiated personal names. With the able and generous assistance of a dozen cooperating libraries, more than 8400 of these records were evaluated and converted manually. In the process, several thousand new unique and non-unique authority records were created.

3. Double conversion

These two headings were checked to make sure that they did not "double-convert":

P'i-hsien (Kiangsu Province, China) converted to Pi Xian (Jiangsu Sheng, China)
T'eng-hsien (Shantung Province, China) converted to Teng Xian (Shandong Sheng, China)

4. Subject headings and subject subdivisions for regions in China

Some of the subject headings for regions in China converted correctly, but others did not. Therefore, all headings on bib records for regions in China were located, evaluated, and corrected when necessary.

EXAMPLES:

651 -0 $a Canton Region (China)… [changed manually to Guangzhou Region (China)]

651 -0 $a Taiyuan Shi Region (China) [changed manually to Taiyuan Region (Shanxi Sheng, China)]

650 -0 … $z Sinkiang Uighur Autonomous Region [changed manually to Xinjiang Uygur Zizhiqu]

650 -0 … $z Tangshan (Hebei Sheng) Region [changed manually to Tangshan Region (Hebei Sheng)]

650 -0 … $z Luoyang (Henan Sheng) Region [changed manually to Luoyang Region (Henan Sheng)]

5. Multi-syllable terms for Chinese jurisdictions

Ten multi-syllable terms for Chinese jurisdictions were to have been joined together by the conversion program when they were identified as being part of a proper name. Some, however, were joined together in other situations. Also, some of the correctly converted terms had to be changed. (For example, T'ai-wan ti ch'ü converted to Taiwan Diqu; this string had to be changed to Taiwan di qu, because the term di qu (地区) in this instance refers to the Taiwan region in general, and not specifically named location.) We scrutinized each bib record on which these ten terms appeared; many records were corrected.

Term	Hits	Needed Correction
diqu	1670	ca. 1100
tequ	88	ca. 55
xingzhengqu	75	ca. 40
zhuanqu	11	2
dujiaqu	1	0
ziran	33	0
zizhiqi	1300	0
zizhiqu	11	0
zizhixian	253	0
zizhizhou	356	0

6. Bogus multi-syllable terms

On Chinese bib records converted in RLIN, the conversion program incorrectly created several multi-syllable generic terms. These are the terms that have been identified and corrected:

Wade-Giles syllables	Converted to	Should be
ti ch‘üan	diquan	di quan
ti ch‘üeh	diqueh	di que
tu chia ch‘ü	dujiaqu	du jia qu
min tsu	minzu	min zu
te ch‘üan	tequan	te quan
hsing cheng ch‘üan	xingzhengquan	hsing cheng quan
tzu chih ch‘üan	zizhiquan	zi zhi quan
chuan ch‘üan	zhuanquan	zhuan quan

7. Guangzhouese

On Chinese bib records converted in RLIN, the word Cantonese was converted to Guangzhouese when it appeared in subject headings. This term has been manually corrected on all LC records.

EXAMPLES:

650 -0 $a Guangzhouese dialects [changed manually to Cantonese dialects]

650 -0 $a Cookery, Chinese $x Guangzhouese style [changed manually to Cookery, Chinese $x Cantonese style]

CLEANUP TASKS THAT ARE CURRENTLY UNDER WAY

1. Chinese serial records that were marked for review in the 987 field

The ca. 800 remaining serial records that were marked for review are being converted.

2. Unconverted access points on non-Chinese serial records

Serial records needing changes to access points will be identified in the course of performing the following cleanup tasks, and sent to serials catalogers for correction.

3. Chronological subdivisions

Chronological subdivisions are being systematically converted from the list of subdivisions that appears on the pinyin home page.

4. Headings for Chinese jurisdictions; conventional place names

Almost all authority records and headings for Chinese jurisdictions on

Chinese bib records were correctly converted by the machine program. Most of the headings on Korean and Japanese records on RLIN have also been converted. Headings for conventional headings for provinces are now being corrected on non-Chinese and PREMARC records. Because of the many recent changes to the names and boundaries of Chinese cities and counties, a comprehensive review of these headings will be conducted at a later time.

5. Subject headings that did not convert, or needed to be changed

Most subject headings on RLIN records have been corrected, with reference to the four lists of Chinese subject headings that appear on the pinyin home page. Subject headings on non-Chinese and PREMARC records are now being converted.

6. "Most frequently used" headings

The "most frequently used" headings in the LC database are being systematically identified and converted on non-Chinese and PREMARC records. To date, about 80 of these headings have been converted, on a total of some 9000 bib records. We estimate that conversion of the remaining 45 headings will require changes on about 5000 more records.

7. Wade-Giles headings on bib records, identified by $wnne and $wnnea references

We plan to extract from files of converted name authority records the former headings, which are coded either $wnne or $wnnea, and then run them against bib records in the LC database to identify headings that need to be converted.

8. Syllable sweep for bib records for instrumental music

Unique Wade-Giles syllables are searched in music records in the LC database. All records that appear to include romanized Chinese are printed out, reviewed, and converted where appropriate. There were 400 music records that included the syllables *chang* and *cheng*, and 127 of them were converted to pinyin.

9. Syllable sweep for bib records for motion pictures

Unique Wade-Giles syllables will be searched in motion picture records in the LC database. All records that appear to include romanized Chinese will be printed out, reviewed, and converted where appropriate. Since much of the data on these records that appears to be romanized has, in fact, been transcribed from copyright applications, titles proper on motion picture

records will usually not be converted. Pinyin data may be added to 246 fields. It is estimated that the searches will result in about 1550 hits, perhaps 250 of which will be changed.

10. Systematic 041 and 043 searches on Voyager (ca. 2500 records)

A series of searches of the 041 and 043 fields will be conducted to identify records that contain unconverted romanized Chinese strings or headings.

11. Chinese monograph records that were marked for review in the 987 field

The term [non-access] is being added in the 987 $f subfield of bib records that have been marked only for change to a non-access point. The remaining marked records will be reviewed. Records on which access points need change will be converted or corrected; the others will be marked [non-access] and set aside.

12. Unconverted IBC serial records

The ca. 600 brief Chinese acquisition records in the LC database will be reviewed and converted.

13. Names of geographical features (rivers, mountains, deserts, etc.):

The conversion program connected certain generic terms for geographic features (primarily the terms for rivers) for geographic features to the names that preceded them. These generic terms will be identified and separated on authority and bib records, to conform to the romanization guidelines.

EXAMPLES:

pre-conversion WG form	machine converted to:	change to:
Chang-chiang	Changjiang	Chang Jiang
Huang-ho	Huanghe	Huang He
Chu-chiang	Zhujiang	Zhu Jiang

At the same time, some 20 multi-syllable generic terms which are used in proper names were not connected by the conversion program. They will be identified and joined together when appropriate.

EXAMPLES:

pre-conversion WG form	machine converted to:	change to:
Huang-t‘u kao yüan	Huangtu gao yuan	Huangtu Gaoyuan

Ch‘ing Tsang kao yüan	**Qing Zang gao yuan**	**Qing Zang Gaoyuan**
San-chiang p‘ing yüan	**Sanjiang ping yuan**	**Sanjiang Pingyuan**
T‘a-k‘e-la-ma-kan sha mo	**Takelamagan sha mo**	**Takelamagan Shamo**
Ch‘ai-ta-mu pen di	**Chaidamu pen di**	**Chaidamu Pendi**
Su-i-shih yün ho	**Suyishi yun he**	**Suyishi Yunhe**
Pa-na-ma yün ho	**Banama yun he**	**Banama Yunhe**

14. min guo → Minguo

When the syllables *min guo* together are used to mean the Republic of China, they must be capitalized and connected. The conversion program did not do this. There are perhaps 500 authority records and many hundreds of bib records that need to be changed.

EXAMPLE:

pre-conversion WG form	**machine converted to:**	**change to:**
Chung-hua min kuo	**Zhonghua min guo**	**Zhonghua Minguo**

15. Tibetan language bib records

Records for Tibetan material will be reviewed because many of them have Chinese colophons and romanized Chinese data in access points.

16. ”Title in Chinese”

A search for the phrase “title in Chinese” calls up 977 hits in the LC database. These records will be reviewed and converted, because most of them are non-Chinese records that include romanized title added entries.

CLEANUP TASKS THAT ARE NOT BEING PURSUED AT THIS TIME, BUT WILL BE EVALUATED AT A LATER DATE

1. Capitalization of generic terms for place names

The conversion program did not capitalize generic terms for place names, as called for by the romanization guidelines. This problem does not affect filing or access. These terms are now being capitalized on an as-encountered basis.

2. di / de

The conversion program automatically converted the syllable ti to di. The

romanization of the character 的, therefore, converted to di rather than de. This syllable is now being changed on an as-encountered basis.

3. Bib records marked for review in non-access points

Bib records that were marked for review because of an unconverted or questionable string of text in a non-access point are being set aside, and may be changed later on. There will probably end up being a total of some 6000 such records.

4. 880 fields

Portions of 880 fields sometimes did not convert, or converted differently from their parallel roman fields. Some of the reasons for this occurrence are explained in the section of the home page that describes the conversion of bibliographic records. These inconsistencies will probably be corrected on an as-encountered basis.