PINYIN STANDARDISATION: THE OPTIONS

The Australian libraries with the largest collections of Chinese-language materials use Wade-Giles romanisation. There are four Australian libraries with substantial collections using Pinyin romanisation, who cannot share in the national database because Wade-Giles romanisation is the standard. As a result of their concern, the Australian Bibliographic Network (ABN) is looking at the implications, for the authority file and retrieval in particular, of running parallel records in Wade-Giles and Pinyin and is seeking comments on Pinyin standardisation.

It is clear that, if libraries using Pinyin are to benefit from shared cataloguing in a network, the first step is standardisation for word division and capitalisation. Punctuation does not appear to be a problem in this instance. In addition, some attention must be given to geographic names in the Publication field.

The definition of what constitutes a "word" is unclear in Chinese. In Chinese script, there is no other division than by character; Chinese characters are basically syllables which may stand alone as what we would call "words" or be part of a "compound word", depending on their meaning. The main difference from their romanisation counterparts is that each character is distinctive and conveys its meaning despite similarities in pronunciation (that is, homophones are not represented by homographs). Here follow some observations based on current practices on ABN, as well as on information from libraries using Pinyin.

WORD DIVISION AND CAPITALISATION ON ABN

Monosyllabic Pinyin word division appears on ABN for an increasing number of records loaded from the British Library. Some Australian libraries are adding holdings to these records, and some are putting in Pinyin records with a variety of word divisions.

Occasional records appear for libraries which are not generally cataloguing Chinese-language material on ABN, such as Macquarie and Monash University Libraries; a Pinyin record has even been entered by the National Library of Australia. The reason for this seems to be that, when the publisher supplies romanisation on the publication, general cataloguing staff decide that they have enough information to catalogue the item themselves, without referring to East Asian cataloguers.

TABLE 1 SAMPLE CHINESE BIBLIOGRAPHIC

NUCOM symbol	Roman-isation	General text	Personal names	Corporate names/ Publisher names
LC	Wade-Giles	Monosyllabic: Chung-kuo hai yang yu yeh chien shih	Hyphenated given names: Ch'en, T'ao-sheng	First word capitalised: Ch'ing kung yeh ch'u pan she
ANL	Wade-Giles	Monosyllabic	Hyphenated given names	First word capitalised
BNB	Pinyin	Monosyllabic: Ying Guo su miao ji shui cai hua zhan	Joined given names	First word capitalised
NMQU	Pinyin	Polysyllabic: Zhongguo zhedueshi yanjiv [sic];	Joined given names	First word capitalised
SFU	Pinyin	Polysyllabic	Joined given names	All words capitalised
VERC	Pinyin	Monosyllabic: Wan qing zhong guo wai xiao hua	Joined given names	(no examples found)
VU	Pinyin	Polysyllabic	Joined given names	First word capitalised
VU	Wade-Giles	Monosyllabic	Joined given names	First word capitalised
WMDU	Pinyin	Polysyllabic, short words separate: Dili xue cidian	Joined given names: Fang, Weizhong	All words capitalised: Zhongguo Shehui Kexue Chubanshe
WU	Pinyin	Monosyllabic	Separate given names: Liang, Wen Sen Chen, Jin Quan	First word capitalised: Shanghai ren min chu ban she
ZNUC	Pinyin	Polysyllabic, varied capitalisation: Zhongguo gongcheng xuekan; Zhongguo Dizhi Kexueuyan[sic] Yuanbao		All words capitalised: Zhongguo Kexueyuan Linye Turang Yanjiusuo

RECORDS ON ABN

NUCOM symbol	Geographic names	Place of Publication	Notes
LC	Hyphenated, first word capitalised	Romanised: Pei-ching English: [Peking]	
ANL	Hyphenated, first word capitalised	English and Romanised	One Pinyin record found; not NLA standard practice
BNB	Separate, varied capitalisation: Ying guo, Ying Guo	Romanised: Beijing English: London	
NMQU	Separate, first word capitalised	Romanised	Not Macquarie's standard practice
SFU	(no examples found)	Romanised	
VERC	Separate, first word capitalised: Zhong guo	Romanised English: Hong Kong	
VU	First word capitalised	English Romanised	
VU	Hyphenated, first word capitalised	Romanised	
WMDU	All words capitalised: Jiangsu Sheng	Romanised: Guangdong	Wade-Giles and Pinyin headings: Wu, Li-fu (WG) Shen, Guangyao
WU	Separate, first word capitalised	Romanised	Wade-Giles and Pinyin headings
ZNUC	All words capitalised	Romanised: Beijing; Zhongqing [sic] English: Taipei	Errors present. Wade-Giles title in variant field with non-standard capitalisation.

Table 1 shows information about word division in records found on ABN. Library of Congress and National Library of Australia records are included for comparison. A summary of findings follows.

General text word division

Syllables are generally grouped together according to meaning, and in most instances follow a similar pattern,

e.g. Zhongguo tushuguan minglu
Zhongguo yaolixue yu dulixue zazhi

Connectives (yu, ji), particles (de) and short words (shi, xue) may be attached as suffixes,

e.g. Zhongguode OR Zhongguo de
Zhongguo shehui jingjishi yanjiu
Zhongguo zhexueshi BUT Zhongguo gudai huobi shi
Zhongguo yaolixue yu dulixue zazhi
Guowai yuyanxue

Personal names

The most common practice is for the family name to stand alone and the given names to be joined (Pinyin) or hyphenated (Wade-Giles). This practice is widespread, whatever the type of word division for other text. The first letter of the family name and the first given name is in upper case. There are isolated instances on ABN of Pinyin given names being separated.

e.g. Wang, Enguang Wu, Renyong Xiue, Renyong
BUT Chen, Jin Quan Liang, Weng Sen

Corporate names

In records with monosyllabic word division, the main differences between Chinese-language records on ABN are in capitalisation. In those with polysyllabic word division, a corporate name containing a geographical name may be joined and capitalised as follows:

e.g.
Wuhandaxue OR Wuhan daxue (Wuhan University)
Zhongguo kexueyuan OR Zhongguo Kexue Yuan (Academia Sinica)
Zhongguo Kexue Yuan Yuyan Yanjiu Suo
Zhongguo Kexueyuan Linye Tujang Yanjiusuo

Corporate names occur in the Publisher field as well as in the Corporate Author and Title fields. Word division and capitalisation should be standardised for each field.

Here are some examples of varying corporate names in the Publication field on ABN:

Taipei, Taiwan : Zhongguo Gongchengshi Xuehui, Taibei.
Beijing : Zhongguo Shehui Kexue Chubanshe, 1984.
Tianjin : Tianjin ke xue ji shu chu ban she, 1984.
Shanghai : Shanghai ren min chu ban she, 1983.

In the Anglo-American Cataloguing Rules 2nd edition (AACR2), Appendix A.34, the rule states: "...If the language has no system of capitalisation, capitalise the first word of a title or a sentence and the first word of the name of a corporate body or a subdivision of a corporate body. Capitalise proper names according to English usage."

Geographic names

Geographic names may occur in headings, in titles or in the Place of Publication subfield. AACR2 has guidelines for geographic names in headings (Chapter 23) and for geographic names in the Publication field.

As far as romanised titles are concerned, if place names are divided by syllable, a decision should be made on capitalisation of each part of the name. These variations were found on ABN:

Zhongguo	Zhong Guo Ying Guo	Zhong guo Ying guo	zhong guo

Both 'Zhong Guo' and 'Zhong guo' were found in British records.

On ABN there is a mixture of English and Chinese in the Place of Publication subfield. In AACR2, Rule 1.4B4 states "Give names of places, persons, or bodies as they appear, omitting accompanying prepositions unless case endings would be affected", and, Rule 1.4C1 (revised version of 1.4B5): "If the name of the place appears in more than one language or script, record the form in the language or script of the title proper. If this criterion does not apply, record the form that appears first."

In other words, the rule says that the place of publication should be treated in the same way as the title proper (e.g., romanised if the title proper is romanised), but, failing that, other forms of the place name occuring on the item are permissible. Library of Congress practice is to romanise the place name or, if it is not present, to supply the English version in square brackets.

Even the English versions of Chinese place names vary: has the word 'Beijing' reached the stage where it has become the commonly accepted English version in preference to 'Peking'? How about other place names such as 'Nanjing'? These are some Chinese geographic names:

English	Chinese	
	Pinyin	Wade-Giles
Canton	Guangdong	Kuang-tung
Chungking	Chongqing	Ch'ung-ch'ing
Hongkong	Xianggang	Hsiang-kang
Nanking	Nanjing	Nan-ching
Peking	Beijing	Pei-ching
Shanghai	Shanghai	Shang-hai
Singapore	Xinjiapo	Hsin-chia-p'o
Tientsin	Tianjin	T'ien-chin

Thus, in addition to a choice of romanisation methods, more than one English version of the place name may also occur, of which one, however, may be identical to the Pinyin version; there are probably no more than three possibilities. The Place of Publication subfield is not an access point in the card catalogue, and at present the subfield is not searchable online; on the other hand, it may become searchable when STAIRS search software is installed on the ABN database. Unless a consistent form of place name is used, items may be difficult to retrieve on Place of Publication.

Parallel titles

Parallel titles are singularly unreliable as a source of information. For a start, they do not always reflect the full Chinese title; the following is not a lone example:

Wuhan daxue xuebao

This was found on a serial when the full Chinese title was in fact:
Wuhan daxue xuebao. Ziran kexue ban.

Secondly, word division varies considerably,

e.g. Yuwenyuekan
(yuwen = 'language', run together as one word)
Yuwen xuexi
(yuwen = 'language', separate word)
Zhongguo jingjiwenti
(jingjiwenti = 'economic problems', run together as one word)
Jingjiyu guanli yanjiu
(yu = 'and', suffixed to preceding word)
Zhongguobiaozhunhua (one long word)

Thirdly, since many Chinese are not at ease with romanisation, there may be a number of errors in transcription.

When such 'parallel titles' are taken as the title proper and copied exactly as they appear on the item, any unusual word division, omissions or errors are reflected in the record; an entry results for a title which is inaccurate and very difficult to retrieve.

Authority work

Some libraries entering Pinyin records are also checking Wade-Giles headings, for example, for personal names. Where a Wade-Giles heading is found, it is attached to the record, but where there is none, a new Pinyin heading is being created. The resulting record contains headings in a mixture of romanisation methods; other cataloguers may not think to check for Pinyin headings, and the authority file is likely to contain two unlinked entries for the same author.

SOME INDIVIDUAL LIBRARY POLICIES

The British Library follows the methods of the printed cards published by the National Library of China with some variation. Each character is filed separately according to Pinyin romanisation. Place names and personal names are joined as seems appropriate; only the first letter of the first word of corporate names is in upper case, as in Wade-Giles. The Library follows, in order of priority: Xin hua zi dian, Ci hai and Zhong hua da ci dian (Taipei), converting the latter from Wade-Giles to Pinyin.

In the EALGRA Newsletter no. 14, Pauline Haldane reported that the **Berlin State Library** uses Pinyin romanisation with two to three romanised characters connected as words with hyphens. The computer then treats the connected cluster as one word. This use of hyphens allows the user to search on compound words; at the same time, the syllables could be treated singly when necessary; for example, a Wade-Giles - Pinyin conversion program could be developed which treats hyphens as spaces. Thus there could be both precision for retrieval and ease of conversion between romanisation methods. At the same time, rules would still be needed for grouping the clusters.

Griffith University Library uses monosyllabic word division for text, corporate and geographic names (e.g. Zhong Guo); personal given names are joined.

Macquarie University Library divides text character by character; where Library of Congress hyphenates two characters together, Macquarie joins them. Geographical and personal names are similarly treated. Macquarie uses LC cards for copy cataloguing and replaces Wade-Giles with Pinyin, sometimes supplying the Wade-Giles version in a note. The main entry is in upper case and the romanised title, when not the main entry, is capitalised according to LC practice for Wade-Giles.

Murdoch University Library joins the words in sentences or titles together according to the sense; whenever there is doubt, they prefer the word to stand alone. Particles and connectives stand alone. Personal names follow the standard practice of joining the two given names, while corporate names are divided according to the above principle, e.g. Zhongguo Kexue Yuan Yuyan Yanjiu Suo. Geographic names are joined except for words such as 'Sheng' (province): Jiangsu Sheng.

A small sample of **The National Library of China's** printed cards shows monosyllabic Pinyin word division and no capitalisation. Given names are joined as author entries, but when a personal name occurs in the title, the given names are separate. Tonal markings are used.

DISCUSSION

There appears to be least divergence over the form and capitalisation of personal names. Corporate and publisher names suffer from the same problems as other sentence or title words, but contain more variations in capitalisation.

The chief source of inconsistent Pinyin romanisation and capitalisation is most definitely the publisher. This is unfortunate, since libraries without a Chinese specialist rely on this type of information. Even libraries with East Asian cataloguers may not pass Chinese-language material to their specialist staff if romanisation is added by the publisher. There is a conflict between AACR2 requirements to put down details as they appear on the publication, and the need for consistent romanisation.

The methods employed by libraries of grouping "words" are similar, but minor variations affect retrieval. Although libraries join syllables as seems "logical", their systems of logic do not necessarily coincide. In Chinese there is a tendency towards balanced pairs of syllables; doubt arises about dealing with extra syllables.

Monosyllabic word division is the solution which has already been adopted by the British Library. Not only does it remove doubt over word division, but it could facilitate automatic conversion

to and from Wade-Giles. However, some precision may be lost in title searches: search software which offers retrieval by title keywords only may retrieve large sets of documents; search software with proximity operators or an exact match option would be required for efficient retrieval.

As noted previously, polysyllabic word division with short words and particles treated as suffixes could present some problems for retrieval; extensive use of truncation may be necessary, resulting in large retrieval sets and lowered precision.

How important is it for capitalisation to be standardised? Does it only matter if it affects retrieval? And hyphenation? In English there is a lack of standardisation in free text terms, for example, 'on-line' and 'online', 'data-base', 'data base' and 'database'. Generally speaking, a user cannot search on punctuation, but punctuation and hyphenation which has been entered in a record may nevertheless affect retrieval. On ABN, a hyphenated word is treated as two words if the second word begins with a letter in upper case and as one word if the second word begins with a letter in lower case. The effect of this is to treat such words as 'on-line' as one word for retrieval purposes, while treating double-barrelled surnames as two words.

Ease of retrieval should be borne in mind when deciding on a method of Pinyin word division. In the past, shared cataloguing databases have been developed on the assumption that the users are librarians who are familiar with Library of Congress subject headings (LCSH), and that they are searching for known items. The search software is not oriented towards performing subject searches on free text in the Title field. Nor is a subject search in the Subject fields always easy on ABN. Yet there may be occasions when reference librarians need to search on concepts in the Title field, if only to find a relevant record and examine its subject headings for further searching. With the arrival of Online Public Access Catalogs (OPACs), inhouse or dial-up, it can no longer be assumed that librarians are the only users.

OPTIONS

1. Adopt monosyllabic word division along the lines of the British records on ABN and develop a comprehensive policy to cover all situations, in consultation with East Asian librarians in Australia, in order to form a national standard for Australian libraries.

2. Adopt polysyllabic word division and develop rules to cover all situations, based on current practices in Australian libraries and in consultation with East Asian librarians in Australia, in order to form a national standard for Australian libraries.

3. Adopt monosyllabic word division but with the addition of hyphens in order to group syllables into compound words, along the lines of the Berlin State Library's practice.

When developing guidelines, it is necessary to bear in mind the conflict between AACR2 rules for transcribing details exactly and the variations in publishers' romanisation and capitalisation for corporate names, etc., also to bear in mind the implications of the final choice of word division for online retrieval, conversion between romanisation methods, and for the wider exchange of bibliographic records.

It is most regrettable that, whatever the final choice of word division, some libraries will be disappointed. None has the resources for a major changeover, and all probably suffer from the "low priority for East Asian cataloguing" syndrome.

CONCLUSION

While the ABN Standards Committee is deliberating whether to allow two romanisation methods on ABN with a resulting duplication of bibliographic records and authority headings, the two romanisation methods are already on the system, and a standard should be adopted for Pinyin as soon as possible.

Some issues and options have been suggested above; there may be others not covered here, and readers are invited to contribute to the discussion. In particular, the needs of the library user should be considered. The author hopes that the resulting exchange of information may be useful to ABN, to libraries with Chinese-language collections who are thinking of joining ABN or networking elsewhere, and to any libraries developing a policy for new Chinese-language collections.

Susan MacDougall
Canberra College of Advanced Education

The author would like to thank colleagues in Australian libraries and the British Library for contributing information for this article.