WADE-GILES AND PINYIN AS ONLINE RETRIEVAL TOOLS

INTRODUCTION

Both Chinese romanisation methods, Hanyu pinyin (Pinyin) and Wade-Giles, have co-existed for some time. Wade-Giles is extensively used in libraries in the English-speaking world, while Pinyin is commonly used for language-teaching. Although both methods present some problems for online retrieval there may be instances when a title search is the only practicable means of access to a record. In this article the two romanisation systems are compared, tested briefly on the Australian Bibliographic Network, and the implications for retrieval discussed with reference to Chinese serials. A number of serials contain both romanisation methods; they tend to have short, general titles consisting of relatively high frequency words. For these reasons, Chinese serials illustrate well the problems of retrieval by romanisation.

ACCESS TO CHINESE BIBLIOGRAPHIC RECORDS

ISSNs have only recently started to appear on some Chinese serials, chiefly scientific ones. The Chinese use two types of serial number which, however, are not used uniformly throughout China and cannot be considered an exact equivalent to ISSNs. For a fuller explanation of Chinese serials numbers see my recent article (1). In addition, they have not commonly been added to online bibliographic records and are not generally available online (except on URICA at Australian National University Library). Therefore it is in only a few cases that there is access to a serial record via a unique number other than the system number.

Choice of other access points may be limited. It is common cataloguing practice among cataloguers of Chinese materials not to enter a corporate author if it is a repeat of the title plus the words 'Editorial Board'. Thus the author fields may not be available for searching. Since the general nature of serial subject headings means that a subject search can retrieve large sets of records, a subject approach is not efficient when one specific title is required. In many cases a title search, which may well be the library user's first choice of access, is in fact the only useful option.

CHARACTERISTICS OF SEARCH SOFTWARE

In general, the search software of library bibliographic utilities is less flexible than that of full information retrieval systems. The search software of the former has no backreferencing, limited use of Boolean logic and no positional operators. Most search software ignores diacritics and special characters but varies as to how it treats punctuation. On the Australian Bibliographic Network (ABN) and the Research Libraries Network (RLIN), a hyphenated word is

treated as one word, while on URICA it is treated as two separate words.

Indexing policies may vary. Some systems do not index words with fewer than three letters. This affects the retrieval of Chinese romanised records, since short words with significance in Wade-Giles romansation, such as 'a', 'an' and 'i', are not indexed. Words in the stopword list which may be significant in Chinese are also inaccessible in a keyword search. A further problem is that the system may not be able to handle high occurrences of words, resulting in suppressed keywords or in messages to the effect that the retrieval set is too large.

URICA offers a direct hit or 'exact match' facility as well as the keyword title search, and RLIN offers a truncatable exact match ('title phrase' in RLIN terminology); this is very useful in reducing the size of the retrieval set in the absence of positional operators, since it means that the order of words in the title is taken into account.

RLIN supplies a further aid to retrieval by putting serials in a separate file from monographs. The files can be searched together or separately.

WADE-GILES, PINYIN AND ONLINE SEARCHING

Tonal languages can be expected to have problems of high recall and low precision; these languages have a high number of homophones and rely to varying degrees on special characters, diacritics or tonal indicators, which are ignored in online searching, to distinguish different words of the same romanised spelling.

Under the Wade-Giles romanisation method, aspirated consonants are differentiated from unaspirated consonants with a special character, the 'ain', which looks like an apostrophe but is not. Since punctuation and special characters are stripped by search software, both the aspirated and unaspirated words will be retrieved when either is searched. An extreme example is the words chu, ch'u, chü and ch'ü; in Mathews dictionary (2) they represent 51, 28, 59, and 34 Chinese characters respectively. When any one of these words is searched online, theoretically at least, all meanings could be retrieved and precision would be low. A personal author search on the same words could retrieve up to 13 different surnames.

Additionally, each syllable in Wade-Giles, which represents one Chinese character, is entered separately even 'though it may be part of a compound word. The exceptions are geographical names and personal given names, which are entered as hyphenated words. The Library of Congress laid down romanisation, capitalisation and punctuation rules in its Cataloging Service Bulletin 118. (3); thus, Wade-Giles word division is standardised.

The Pinyin romanisation method does not have the problem of diacritics or special characters (Wade-Giles spellings chu, ch'u, chü and ch'ü are differentiated in Pinyin by being spelt zhu, chu, ju and qu respectively), but it does have a problem with word division. Although a generally accepted word division method exists, it is not

followed universally, and not by serial producers in China. Where Pinyin romanisation is included on the chief source of information (title page or cover) as a parallel title, the 'words' are divided in different ways, varying even from issue to issue of the same serial title. This may present the cataloguer with a dilemma, given the AACR2 (4) rule that wording in the catalogue record should be exactly the same as that on the title page. Not only is the romanisation on the title page not always an exact equivalent of the title in Chinese characters, but it is sometimes used as a decoration of the title page and repeated across it.

ONLINE RETRIEVAL RATES

Results of a search on ABN of 300 titles of varying lengths, taken from ANU Library's 1986 Chinese serial list, are set out below. The list contained very few titles in certain categories; these are indicated.

AUSTRALIAN BIBLIOGRAPHIC NETWORK

Average retrieval set (number of records)

Title length	2 words	3 words	4 words	5 words	6 words
Wade-Giles one-syllable words	232.5	20.5	19.5	6.4	.8
one two-syllable word in title	*66	10.2	9.28	.78	nf
Pinyin two-syllable words	.5	*.4	nf	nf	nf

*small sample set nf = no figures

As there were far fewer Pinyin titles than Wade-Giles titles on the database, figures were not directly comparable. Retrieval rates depended on how many Chinese characters were represented by the romanised word as well as the frequency of each; bisyllabic words reduced the number of false drops considerably and longer titles reduced the size of retrieval sets.

Wade-Giles was more reliable because there was rarely any doubt as to how to divide the words. It was quite likely that, barring inputting errors, all titles containing the search words were found, relevant or not. Retrieval rates showed considerable variation; the figures for titles of two monosyllabic words ranged from 0 to 2195 records. Short titles were the most difficult to retrieve since a great deal of browsing was needed. One ABN screen could display up to 10 records (via dial-up access); very large sets involved calling up successive screen displays without any guarrantee that the required title was in the set. Unless the searcher had considerable time and patience, those titles were as good as lost.

For Pinyin, on the other hand, recall was relatively low and titles were missed because of variations in word division. When a title was found, it was usually relevant and the only one in the retrieval set. Most records held the Pinyin title transcribed exactly in the parallel title field and again in the variant title field. A few Library of Congress records also held a standardised version of the parallel title in the key title field. Such records could be retrieved both ways. This appeared to be because of variations in Pinyin word division between issues of the same serial rather than an effort on the part of Library of Congress to address the problem of Pinyin word division.

CONCLUSION

The investigation revealed problems caused by the very nature of tonal languages, by romanisation methods and by unsophisticated search software. For reasons discussed, it is apparent that, in a keyword search, searching on Wade-Giles romanisation generally produces larger retrieval sets than Pinyin, but that variations in Pinyin word division detract from its usefulness as a retrieval tool. Without standardisation, the same search must be done more than once to cover the possible variations, wasting valuable time. With standardisation, Pinyin would be the more efficient tool.

There are already retrieval difficulties on ABN caused by the nature of short serial titles searchable only as keywords. Although not tested here, URICA's facility for the exact match and RLIN's facility for the truncatable exact match appear to be a distinct advantage. As the size of the database grows, the size of retrieval sets will increase proportionally. Unless the software is enhanced to keep pace, the problems will increase.

Many libraries in the English-speaking world rely on Library of Congress for copy cataloguing; they are unlikely to make cataloguing policy decisions unless they are prepared to edit every downloaded record. Therefore they look to Library of Congress to take the lead; if Library of Congress is not already addressing the matter of Pinyin access to records, it should do so.

Susan MacDougall
Canberra College of Advanced Education

NOTES

(1) MacDougall, S. "Chinese serial numbers: an aid to serials management." EALRGA Newsletter no. 10. (December 1986): pp 44-45.

(2) Mathews' Chinese-English dictionary. revised American ed. Cambridge, Mass: Harvard University Press, 1960. pp 186-232.

(3) Library of Congress. Cataloging Service Bulletin 118 (Summer 1976): pp 35-55.

(4) Anglo-American cataloging rules. 2nd ed. London: LA, 1978.