Libraries Australia – improvements to Chinese and Japanese language searching

Di Pin Ouyang【歐陽迪頻】
Asian Collections, National Library of Australia

Recently, Libraries Australia implemented changes that have greatly improved the searching of Chinese and Japanese script characters. Libraries Australia users can now do accurate simultaneous searching or browsing using script characters and be confident they have retrieved all of the relevant records in a single search. This has significantly improved the efficiency of our business in the Chinese and Japanese Units of the National Library. It is a great improvement and makes the script searching functionality much more useful!

Chinese characters are called Hanzi, which can be divided into two groups, Traditional Chinese (used in Taiwan, Hong Kong and Macau) and Simplified Chinese (used in mainland China). Previously, Libraries Australia records that included either of these character sets had to be searched separately. We had to constantly switch between simplified and traditional characters in order to execute a comprehensive search across all materials available in the National Bibliographic Database (ANBD), in effect conducting a double search every time script characters are involved.

A searching problem also existed for Japanese characters. In Japanese, a word may consist of several characters to convey meaning and there are no spaces between these characters. However, the Libraries Australia search broke up every Japanese word into individual characters which made searching very difficult.

For Chinese, there is now simultaneous searching of both traditional and simplified characters. For Japanese, the updated system is treating consecutive characters as one word even when they are made up of a combination of different types of characters such as *kanji* (= Chinese characters used in Japanese) and either of the two sets of phonetic alphabets (*hiragana* and *katakana*). Queries can include any combination of scripts as long as they are from meaningful character sets.