OCLC CJK Automated Library Information Network

*Andrew H. Wang**

INTRODUCTION

OCLC Online Computer Library Center operates an international computer and telecommunication network. Through this world's largest bibliographic network, libraries acquire and catalog library materials, arrange interlibrary loans, maintain location information on library materials, retrospectively convert card file into machine-readable records in support of library automation, gain access to other databases, order custom-printed catalog cards and machine-readable records for local catalogs.

OCLC established and pursues two goals for its membership:

1. To reduce the rate-of-rise of per-unit costs in libraries; and
2. To increase the availability of library resources to library patrons.

Over 6,000 libraries in all of the 50 states of the U.S.A. and in Australia, Belgium, Canada, China, Denmark, Finland, France, Ireland, Saudi Arabia, Sweden, Switzerland and the United Kingdom use the OCLC Online System. About 25 to 35 new users join OCLC each month. OCLC's membership includes many kinds of libraries: academic, research, school, public, federal, state, municipal, corporate, law, medical, and theological, etc.

---

* Andrew H. Wang, Manager, Online System Products & Services Department, OCLC, U.S.A.

The hub of the OCLC Network is the Online Union Catalog, a database of over 13 million unique bibliographic records, the world's largest database of its kind. This database grows by about 30,000 records weekly, or about 1.5 million records annually. In addition to the bibliographic records, the Online Union Catalog also maintains over 210 million library location symbols.

Libraries can access the Online Union Catalog through either a dedicated line or dial access. Dial access users can dial through Telenet or TYMNET, or dial directly to OCLC Online System. OCLC Network supported 7,618 dedicated-line terminals and 1,784 dial access authorizations as of January 31, 1986 and received an average of 42.22 telecommunication messages per second, a total of 54 million transactions in the month of January, 1986.

Automations of the Chinese, Japanese, and Korean Characters

Since its initial operation in August, 1971, OCLC Online System has been able to process information in roman alphabets and arabic numerals only. Information in nonroman alphabets must be romanized before it can be processed by the OCLC Online System. Unfortunately, romanized Chinese, Japanese, and Korean (CJK) characters are very difficult, if not impossible, for readers of these languages to understand. As a result, librarians who handle materials in these languages have seldom used OCLC Online System for cataloging, interlibrary loan, serials control, serials union listing, and acquisitions purposes. Despite its riches in information, the Online Union Catalog has been of limited use to scholars and students who need materials in these Asian languages. However, this situation is about to change.

Recognizing the needs of librarians, scholars, as well as other library patrons of CJK materials, OCLC made a public announcement and a commitment on 1983 October 18 to automate the CJK characters. At the onset of the CJK project, OCLC established the following five fundamental guidelines for developing the CJK project.

1. Hardware must be a multi-purpose workstation, instead of being a dedicated workstation limited to CJK usage only.
2. Keyboard must be an English-language keyboard so that extensive training on operation of keyboard will not be necessary.
3. Various input methods must be provided for both dedicated CJK staff as well as casual CJK users.
4. Software must be designed in a modular fashion so that additional capabilities can be added without having to redesign the entire software or workstation.
5. Price must be resonable and affordable.

CJK Workstation

OCLC's CJK is an enhancement to the cataloging function of OCLC Online System. The only hardware required for CJK operation is the front-end CJK Workstation. OCLC CJK Workstation is built on OCLC M300 Workstation with 640K memory, supporting American Library Association (ALA) characters.

OCLC CJK Workstation is capable of accessing OCLC Online System through either a dedicated line in synchronous mode or dial access in asynchronous mode. Existing OCLC M300 Workstations are upgradable to OCLC CJK Workstations.

Input Methods and Character Subsets

Since there is no one single best input method for all users, OCLC CJK provides the following five input methods for users of various backgrounds.

1. Tsang-chieh, the only character-based input method of the five, is rated the best input method by Institute for Information Industry. This input method, primarily to be used by dedicated users, can be used for generating Chinese characters in both full and simplified forms, Japanese kanji, and Korean hancha.

2. Wade-Giles, a pronuncuation-based input method, can be used for generating Chinese characters in both full and simplified forms. Since libraries in the United States have adopted Wade-Giles romanization scheme for cataloging Chinese-language materials for decades, catalogers as well as library patrons in the United States are used to this scheme. OCLC adopts the Wade-Giles romanizations, with modifications, provided in the Chinese Character Code for Information Interchange (CCCII) tapes that OCLC obtained from Taiwan.

3. Pin-yin, another pronunciation-based input method, can be used for generating Chinese characters in both full and simplified forms.

4. Modified Hepburn, a pronunciation-based input method, can be used for generating Japanese kanji, katakana, and hiragana. Pronunciation of Japanese kanji varies a great deal based on the context. Due to the time constraint, OCLC can only provide one or two most frequent pronunciations for each kanji. More comprehensive pronunciation will be provided in the future enhancements.

5. McCune-Reischauer, a pronunciation-based input method, can be used for generating Korean hancha and hangul. Since there is no existing dictionary that provides McCune-Reischauer romanization for hancha and/or hangul, OCLC's McCune-Reischauer romanization dictionary will be the first such dictionary to be published.

The following chart summarizes the input methods provided, and the characters they generate.

	Chinese		Japanese			Korean	
	Full Character	Simplified Character	Kanji	Katakana	Hiragana	Hancha	Hangul
Tsang-chieh	•	•	•			•	
Pin-yin	•	•					
Wade-Giles	•	•					
Modified Hepburn			•	•	•		
McCune-Reischauer						•	•

Homophones

All phonetic-based input methods will inevitably generate homophones. OCLC's CJK provides two qualifiers at the end of user's keystrokes to reduce the number of homophones. Users of Wade-Giles and Pin-yin input methods may add to the end of input keystrokes a numeral 1, 2, 3, or 4 to signify the tone of the character desired. In addition, all input methods other than Tsang-chieh may add to the end of input keystrokes as a qualifier, the first letter of Tsang-chieh input method for the character desired.

When homophones occur, the CJK Workstation will beep and display up to eight characters at a time in the homophone block of the status line. Each character so displayed will always be designated by an arabic numeral from 1 up to 8. The user then enters an arabic numeral to generate the desired character. When there are more than eight homophones, the last character in the homophone block will be an up-arrow symbol to indicate

that there are more homophones. The user will depress the space bar to see the next set of up to eight homophones. The last homophone in the list will be followed by a blank space and an exclamation mark.

Screen Display

The OCLC M300 Workstation displays a maximum of 24 lines per screen. Data from the Online System is transferred to the workstation one screen at a time as requested by the user. However, the OCLC CJK Workstation can only display up to 16 lines at a time. Therefore, instead of transferring to the workstation one screen at a time as is the case with the OCLC M300 Workstation, the Online System will transfer all screens of the record to the OCLC CJK Workstation. The CJK Workstation will then store the entire record, and reformat the text for display on the screen. All CJK characters will be displayed in a 16 x 16 dot matrix.

Storage and Retrieval

OCLC will continue to maintain only one database, the Online Union Catalog, for both roman- and nonroman-alphabet records. At the present time, OCLC provides seven possible numeric search keys and four possible derived search keys, as listed below, for searching and retrieving roman-alphabet records in the Online Union Catalog. All of these search keys are also applicable for retrieving CJK records. In addition, CJK users can also use four additional derived search keys in CJK characters.

A. Numeric search keys

1. LCCN: The Library of Congress Card Number.

2. ISBN: The International Standard Book Number.
3. ISSN: The International Standard Serial Number.
4. CODEN: CODENS are five-letter codes assigned to serials by Chemical Abstracts Service.
5. OCLC Control Number: The OCLC control number is a unique number assigned by the OCLC Online System to each bibliographic record as it enters the Online Union Catalog.
6. Government Document Number.
7. Music Publisher Number: Music publisher numbers are plate and publishers' numbers for printed music, and serial numbers and matrix numbers for sound recordings.

B. Derived search keys in roman alphabets

1. Title search key (3, 2, 2, 1) consists of the first three letters of the first word in the title, excluding an initial article, followed by a comma, the first one or two letters of the second word in the title, another comma, the first one or two letters of the third word in the title, another comma, and the first letter of the fourth word in the title.

2. Personal name search key (4, 3, 1) consists of the first four letters of the author's surname, followed by a comma, the first one, two, or three letters of the author's forename, another comma, the author's middle initial which is optional.

3. Corporate name search key (= 4, 3, 1) consists of an equal sign, followed by the first four letters of the first significant word in the name, excluding words on the stop list, a comma, the first one, two, or three letters of the word following the first significant word, another comma, and the first letter of the next word which is

optional.

4. Name/Title search key (4, 4) consists of the first three or four letters of the first word in the author's surname, corporate name, or uniform title, followed by a comma, and the first three or four letters of the first word in the title excluding initial articles.

The user can qualify all derived search keys by type of material, year of publication, and whether the material is a microform reproduction.

C. Derived search keys in CJK characters

1. Title search key (ti:5) begins with a prefix "ti:", followed by the initial one, two, three, four, or five CJK characters in the title. The more characters there are in the search key, the fewer records it will retrieve.

2. Personal name search key (pn:4) begins with a prefix "pn:", followed by the initial one, two, three, or four characters in the personal name, beginning with the surname.

3. Corporate name search key (cn:4) begins with a prefix "cn:", followed by the initial one, two, three, or four characters in the corporate name.

4. Name/Title search key (nt: 1,4) begins with a prefix "nt:", followed by none, or the first character in the name, a comma, and the initial one, two, three, or four characters in the title.

Derived search keys in CJK characters will only retrieve records that contain CJK characters. When variant forms of a character exist, one form of the character in the search key will retrieve records containing all variant forms of that character if these records are retrievable by that search key.

When one search key retrieves more than one record, these records will be sorted by the romanized characters instead of by the CJK characters.

In addition to these 15 search keys, search by subject heading and classification number will be possible in latter part of 1986. Bibliographers and scholars will then be able to compile bibliography by subject area.

Character Set and Interchange Code

Research Library Information Network (RLIN) pioneered CJK project in 1983, and developed RLIN East Asian Character Code (REACC). In order to make OCLC's CJK records compatible with RLIN's CJK records so that both sets of CJK records can be exchanged without technical difficulty, OCLC adopts REACC and its character set. REACC adopts the structure of Chinese Character Code for Information Interchange (CCCII) and contain the following interchange codes:

1. Chinese Character Code for Information Interchange (CCCII)
2. Code of Chinese Graphic Character Set for Information Interchange (CCGCSII)
3. Japanese Industrial Standard (JIS)
4. Korean Information Processing System (KIPS)

RLIN CJK character set consists of 15,850 characters which include 13,650 Chinese characters, 174 Japanese kana, and 2,026 Korean hangul.

Building a CJK Database

OCLC's CJK automation project will begin its field test

phase in April, 1986, and will become available to general users in August, 1986. It is beyond any doubt that the number of CJK records in the Online Union Catalog will grow rapidly thereafter. Following are primary potential sources of CJK records that will enter into the Online Union Catalog:

1. CJK records distributed by the Library of Congress.
2. CJK records from RLIN.
3. Chinese records from National Central Library, and other libraries in Asian/Pacific region.
4. CJK records input or enhanced by OCLC CJK users.

OCLC CJK Software

OCLC CJK software consists of the following three packages.

1. CJK Online Cataloging Package enables users to enter CJK records into, and retrieve CJK records from the Online Union Catalog.
2. CJK Card Production Package enables users to print catalog cards in CJK characters at a local site. CJK characters are printed in 24 x 24 dot matrix.
3. CJK Word Processing Package combines Chinese, Japanese, and Korean language word processors into one package.

Cultural Exchange through Resource Sharing

Technology is of no value to human race unless it provides meaningful service to humanities. The purposes of OCLC CJK Automated Library Information Network are to reduce library's costs of processing materials in CJK languages, to reduce library

staff's time to process these materials, and to increase the availability of bibliographic and location information of materials in CJK languages to librarians, scholars, and students worldwide.

OCLC's CJK database will not only be national in scope, but also international in perspective. It will contain bibliographic and location information of many significant CJK collections in the world. Through access to this database, scholars can compile comprehensive bibliographies by subject, subgrouped by, for instance, dynasty, date of publication, and/or language of materials, etc.

Scholars as well as librarians have had difficult time knowing holding information of any significant CJK library other than that of their own library. In other words, there is no easy way to know what are available, and where to find them. However, through OCLC CJK Automated Library Information Network, this information will be readily available at each scholar's and librarian's fingertips. After National Central Library's records have entered into OCLC's Online Union Catalog, for instance, the bibliographic and location information of National Central Library's collection will become readily available to librarians and scholars worldwide. Knowing exact holdings of National Central Library, scholars in the United States and Europe, for instance, may borrow needed materials from National Central Library, or spend several months in Taiwan to make use of those needed materials for research purpose. OCLC CJK Automated Library Information Network will not only link the scholars and the needed materials, but also provide a shortcut for the East to meet the West.