The East Asian Project of the Research Libraries Group

Alan Tucker
Introduction

In the Spring of next year the Library of Congress, with the largest holdings of East Asian materials in this country, and far and away the largest producer of original bibliographic data on East Asian materials, will take delivery of the first eight of a scheduled total of 24 RLG CJK (Chinese, Japanese, Korean) terminals and begin cataloging online into RLIN, the Research Libraries Information Network. At the same time, the East Asian library of the Hoover Institution at Stanford will receive its four RLG CJK terminals and begin cataloging in RLIN. By the end of 1983, a total of 13 institutions, with 54 RLG CJK terminals, will be using RLIN for the creation, storage, and retrieval of bibliographic records containing Chinese, Japanese, and Korean vernacular data. The installation of RLG CJK terminals in these institutions will mark the fruition of a project to which the Research Libraries Group committed itself in 1979, following the reports of two committees of the American Council of Learned Societies on strategic national planning in the East Asian Library field. The real effect of this is more than simply technical: in using RLIN for technical processing and public services, East Asian libraries will become much more integrated into the processing and service context of their parent institutions, just as their bibliographic records are integrated into the main RLIN bibliographic database of more than 7.5 million records.

There are three components to RLG's East Asian project: the development of the RLG CJK terminal, the development of a character set, and the enhancement of existing RLIN software and the design of modifications to the US MARC format.

The RLG CJK Terminal

In February 1981, RLG circulated a Request for Proposal (RFP) to nearly three dozen firms in the United States, Japan, and the Republic of China, for the development of a terminal which would support Chinese, Japanese, and Korean scripts as well as the complete ALA/MARC character set, be programmed to emulate our existing intelligent terminals, and use our network communications protocol.
Eight potential vendors responded to our RFP, their proposals covering the full spectrum of input techniques from large keyboard to phonetic conversion to component entry. None of the proposed systems was in a state of "off-the-shelf" readiness for our purposes, and after very thorough investigation we entered into a joint development agreement with Transtech International Corporation of Natick, Massachusetts, for the development of what has come to be known as the RLG CJK terminal, as well as for the development of our initial character set, which I will describe later.

The RLG CJK terminal is based on an existing Transtech product, their "Sinoterm," which is already operational in Taiwan. The Sinoterm is a component-entry device, utilizing 245 "word-roots"—complete Chinese characters or basic strokes or shapes—to compose Chinese characters. Unlike some component-entry systems, the Sinoterm does not use component expressions to generate characters in real time; rather, a sequence of keystrokes is used to generate a disk address, and the indicated character is then retrieved from the attached disk and displayed.

Although not the only deciding factor, this approach to character generation was very significant in our selecting Transtech from the other "finalists", as all the proposals which merited serious consideration were based on component entry. RLIN terminals communicate with the host computer at Stanford (a dedicated 16mb IBM 3081) via a polled synchronous network. In normal use, an entire screen is transmitted to or from the terminal at a time. With component entry systems capable of generating spurious characters, character validation would have to take place at the host after the user had typed in a complete screen, perhaps the entire bibliographic portion of a record. There are two problems with this technique: first, since the entire character set could probably not be core-resident, the editing of a screen of data would be heavily I/O consumptive, increasing the number of I/O operations per record by more than an order of magnitude; second, from the ergonomic viewpoint, the terminal operator's efficiency is best maintained by "instant gratification", i.e., if a character is correctly composed, that should be instantly evident, while if a character is incorrectly keyed in, the operator should be immediately notified and should immediately correct the error, rather than having to deal with a complete screen possibly containing a number of incorrect characters.

The Sinoterm approach, of course, neatly solves this problem. It is still possible to key in a character other than the one intended—a given keystroke sequence might simply identify another character—but it is not possible to key in an invalid character, one which does not exist in written Chinese, Japanese, or Korean.

The modifications required to transform the Sinoterm into the RLG CJK terminal were numerous. The existing processor was replaced with a larger one—a DEC LSI 11/23—to handle the extended functionality required of RLIN terminals, and also to change the Sinoterm architecture fundamentally: what had been a single terminal with its own processor and a floppy disk drive for the character set became a terminal cluster, with the processor acting as a cluster controller for as many as four terminals, and with a larger, faster, hard disk (a Winchester) attached to the controller and serving all the terminals connected to that controller.
A larger monitor was used as the display device, in order to allow the standard RLIN screen of 25 lines of 80 alphanumeric characters to be used without modification (pre-formatted input and display screens are an integral part of RLIN); the new monitor also increased the number of Chinese characters which could be displayed on a single line to 40 from the original Sinoterm's 22.

And the keyboard was enlarged and modified. Extra keys were added to support all the keyboard functions available on other RLIN terminals; the full alphanumeric repertoire used on other RLIN terminals was added (the Sinoterm had only an upper-case roman character set); a special row of keys for ISBD-required punctuation was added, duplicating those characters but making them available without regard for the character set selected (about which more below); a kana keyboard layout, based on the Japanese Industrial Standard, was overlaid on the ASCII section of the keyboard; a set of 33 basic hangul was overlaid on the top three rows of the keyboard; and a set of four character-set selector keys was added to enable the processor to interpret the depression of a given key (which might have as many as six characters from three different character sets associated with it) correctly.

In using the RLG CJK terminal for record creation, an operator will always use at least two of the four character sets. Working in Chinese, the user will switch between the roman and Chinese sets for two reasons: because of the Anglo-American cataloging rules, some fields must be entered in the vernacular and others in romanized form or in English; and within fields, at least the content designation—tags, indicators, subfield codes—must be input using the roman set. In Japanese, the worst case, the operator will normally use roman, Chinese (i.e., kanji), and kana (only katakana are engraved on the keytops, but when the keyboard is in lower-case shift, hiragana are displayed, while in upper-case, katakana are displayed). And in some cases, the operator working in Korean will have to make as much use of the character-set selectors as in Japanese, changing from roman to Chinese (i.e., hanja) to hangul.

Roman characters and kana are keyed in on a one-for-one basis, as on a normal alphanumeric keyboard, but both Chinese characters and hangul require multi-stroke entry. The bottom line of the screen on RLIN terminals is used for control and status information; on the RLG CJK terminal, part of that line is used as a "scratch pad" in which the Chinese character or hangul can be composed before being inserted at the current cursor position on the screen. When the operator presses the Chinese (or hangul) character set selector, the scratch pad is cleared and a cursor appears in the first position (the cursor on the main screen is unaffected). As each character component is keyed in, it appears in the scratch pad—Chinese character take from one to eight keystrokes, the average being 3.7, and hangul take from two to four keystrokes. When all the necessary components have been entered, the operator hits the space bar, and the processor interprets this as an instruction to search the disk for a character with a matching keystroke sequence. If such a match is found, the dot-matrix and three-byte codes are retrieved from the disk, the character is displayed at the current cursor position on the main screen, and the code is entered in the data stream within the terminal memory to await transmission to the host when the current transaction segment is complete.
As I mentioned earlier, the displayed character might not be the one the operator intended to input. Every character known to the system has a unique keystroke sequence (with one special class of exceptions which I will describe presently), but if the operator does not mentally decompose the character correctly, another character might be retrieved. A not-very-satisfactory analogy might be the case of someone's typing "quite" instead of "quiet"—all the pieces are there, but arranged differently.

In the event that there is no matching keystroke sequence in the disk file, the operator gets both an audible response and a message in the scratch pad. Because of the way the disk file—which we call the local dictionary—is organized, we expect to be able to indicate to the operator the point in the sequence of input components beyond which no similar strings of components exists.

The exception to the "unique keystroke sequence" rule is a group of roughly 950 traditional-simplified pairs of Chinese characters, based on 18 radicals, which share the same sequence between the members of each pair. To discriminate between the traditional and simplified form, the operator will press the shift key just before pressing the space bar to signify the end of the input sequence; the processor will retrieve the traditional form if the input is unshifted, the simplified form if shift has been pressed. (Incidentally, copious and detailed documentation is being prepared to guide users of the RLG CJK terminal through the many rules and exceptions, regularities, and irregularities, that they will have to handle.)

The Character Set

No one knows with absolute certainty how many Chinese characters there are; no one knows—period—how many are required for the purposes of bibliographic description. Within this context of uncertainty, at least four agencies in different East Asian countries are working on the development of standard character sets for use in data processing. Almost by definition, none of these can ever be totally comprehensive unless a process of continual updating is to be undertaken. For various and unsurprising reasons, the coding systems employed are incompatible, so that a given character may be represented by as many as four different codes. And, for perfectly sound lexicographical reasons, the characters in one set, that developed in Japan for example, may not all be useful or even acceptable in the parent country of another set, say the PRC.

Within RLG, several premises underlay the development of the RLIN East Asian Character Code (REACC) and the basic character set which it represents. The first, and perhaps most important, was that the development of our system for supporting CJK records could not wait for the resolution of the problems outlined in the preceding paragraph. The second was that we should be able to communicate bibliographic data, on tape at least, with the national bibliographic agencies of any East Asian country able to supply or willing to accept such data. And finally, in order to support the accurate transcription of bibliographic data from the materials being cataloged, we needed an expandable character set the initial scope of which should be as broad as was technically and economically feasible.
When we entered into our agreement with Transtech, the existing Sinoterm had a repertoire of 10,934 characters, selected on the basis of the needs and experience of Transtech's Taiwan customers. To satisfy the second and third premises outlined above, we instructed Transtech to carry out a process of conformance between that "native" Sinoterm set and the four existing national standard sets. The results of that conformance are as follows:

-- all characters in volume 1 of CCCII (4,807 characters) were found to be in the Sinoterm set; CCCII codes were added to the file for those characters

-- all but 537 characters in the Sinoterm set were found in CCCII (all published volumes); CCCII codes were added to the file for those characters found (further work is being done on this group of characters)

-- 2,337 characters from the Code of Chinese Graphic Character Set for Information Interchange (i.e., the PRC First Set, 6,763 characters in all) were not found in the Sinoterm set, and were added to it; all characters in this set had PRC codes added to the file

-- 785 characters from the Japan Industrial Standard set (JIS levels 1 and 2 plus extensions -- 6,349 characters) were not found in the Sinoterm set, and were added to it; all characters in this set had JIS codes added to the file

-- 7 characters from the Korean Information Processing System set (KIPS -- 2,392 characters) were not found in the Sinoterm set, and were added to it; all characters in this set had KIPS codes added to the file.

The conformance process yielded a set of 14,063 Chinese characters, for many of which two, three, or even all four of the national standard codes are present. Our internal coding system, REACC, which we will use in records stored in the database and in records distributed on tape to the Library of Congress for their subsequent redistribution, uses the CCCII pattern of coding. There are several reasons for this. First, CCCII currently provides the largest character set available for data processing purposes; by the end of 1982 we expect to have an up-to-date CCCII file containing more than 32,000 characters as a resource for RLIN users. Second, the three-byte structure of CCCII allows for the character set to grow ultimately to include all known Chinese characters. And finally, the internal logic of CCCII makes possible the linkage of traditional, simplified, and other variant forms of characters (and also makes possible some rather powerful indexing and searching, which I will describe shortly).

In our system, characters which do not have CCCII codes use their own national standard codes, but such codes are prefixed by a byte which, in essence, assigns them to one or another of the vacant "planes" within the three-dimensional coding structure reserved for just such other standard (or user-defined) sets.

In referring to the complete CCCII file as a resource, I should explain that there are really two character-set files within the RLIN system. One, partially described earlier, is the local dictionary. This is the set of characters - initially
14,063 Chinese characters and 1,907 hangul -- which is resident on the disk attached to each cluster controller. For each character in the local dictionary, the information recorded consists of the unique keystroke sequence identifying that character (internally, even the members of a traditional/simplified pair have unique keystroke sequences -- the keyboard operator doesn't know this, and doesn't need to (and almost certainly doesn't want to!)), the 16 x 18 dot-matrix representation of the character, and the REACC value for the character. It is this file that the terminal operator is interacting with when keying in characters, and which the online system uses to display records retrieved from the database.

The other file, which we call the CJK Thesaurus, contains all the characters in the local dictionary, and then some. At the outset, it will consist of the complete CCCII set, plus any PRC, JIS, and KIPS characters not in that set, and will thus number well over 33,000 characters. This file is stored as a separate database on the host computer at Stanford, and its principal function is to provide a resource from which characters can be added to the local dictionaries on an as-needed basis. It is the thesaurus which holds the multiple codes for each character, as well as readings in different languages, dot-matrices in different sizes, radical/stroke identification, authentication information (e.g., dictionary references), notes made by the Thesaurus Administrator, and status information, such as whether or not the character has been distributed to the local dictionaries, a rather useful piece of information if the total re-loading of the dictionaries should ever be required.

The Thesaurus Administrator is a member of the staff of the Library of Congress, and is the only person outside RLG central staff who can add to or modify the thesaurus. Users of the system can search the thesaurus by any of its many indexes, for such purposes as determining the keystroke sequence of a character they can't otherwise figure out how to input, or to find a character which isn't available at the terminal, but is needed for the cataloging of a particular work.

In this latter case, the Thesaurus Administrator will carry out a number of tasks to make the required character available. If the character is already in the thesaurus, this will consist mainly of producing the 16 x 18 dot matrix which the RLG CJK terminal uses, possibly adding some codes, modifying some status information, and setting a flag so that the system will be able to extract the character during the next update of local dictionaries. If the character is not in the thesaurus, the administrator will do whatever research is necessary to authenticate the character, then build a complete thesaurus record and ready it for distribution.

Character distribution will take place at night, when the system is not in use, on a cycle which has not yet been determined. With an initial dictionary of more than 14,000 characters, we do not expect the need for additional characters to arise all that frequently, perhaps 100 times a year. As each user logs on, the version number of the local dictionary will be checked against that of the thesaurus. Any discrepancy will prevent the user from logging on; for such a situation to exist, something must have gone wrong with the updating process, and we would prefer to investigate and rectify it rather than have terminals on the network which cannot fully communicate with each other and with the database.
RLIN Enhancements and MARC Modifications

The RLIN Integrated Technical Processing System supports the acquisition and cataloging of materials in all current MARC formats (except Manuscripts), and provides a sophisticated retrieval capability by means of nearly 30 indexes to the bibliographic database. This system is being modified to allow for the creation, maintenance, retrieval, and display of records containing CJK data, and also to make possible the use of such records by users at standard RLIN terminals.

Until fairly recently, the US MARC formats did not have a mechanism by which non-roman character sets could be entered in bibliographic records, nor any adequate way of linking fields which constituted alternate graphic representations of the same information, such as the vernacular and romanized versions of a name. The first of these shortcomings was rectified by the acceptance of a proposal (now being re-negotiated, but for reasons unrelated to this discussion) to define a "character sets present" field, in which one would record the escape sequence(s) used in a record to identify character sets other than the default character set of the agency or country in which the record was created; the proposal covered both single- and multi-byte character sets.

The second shortcoming had to be removed to allow cataloging agencies to produce records which are usable in a heterogeneous environment. In other words, a bibliographic record produced in Japan for use only within Japan need not carry any romanized equivalents of fields entered in kanji and kana. But if, as in RLIN, CJK records are to be integrated into a general bibliographic database (and we in RLG have absolutely no doubt about the correctness of this approach) where they may be retrieved by users at non-CJK terminals, then the parallel entry of at least some basic fields is a necessity. Moreover, since cataloging according to AACR2 requires that access points for names normally recorded in non-roman scripts be recorded using the romanization scheme of the cataloging agency, catalogers using RLIN will continue this practice, while also entering the vernacular form of the access points in order to make possible more efficient searching and retrieval. (One of the basic premises of this exercise, after all, is that romanization is ambiguous!)

To make this possible, we defined a new set of fields within the MARC formats to hold the "alternate graphic representation" of other fields within a record. In some cases, the alternate field will contain the romanized form (of the imprint statement, for example); in others, the alternate field will contain the vernacular form (of the author's name, for example). Each field will carry complete content designation, thus making it properly manipulable, and each will be linked to its associated field in an unambiguous and machine-verifiable way.

This proposed change to the formats was presented to MARBI (the interdivisional committee of the American Library Association responsible for the review of proposed MARC format changes) at its October 1982 meeting, and approved by that group; this will be reflected in a MARC Update to be published by the Library of Congress early next year. At some point in the near future, similar modifications will need to be made to the MARC Authorities format, if control of vernacular headings is to be established; the same technique may be applicable, but I have only just begun to study this.
Given these two extensions to the MARC formats for bibliographic data, the RLIN system enhancements can be described. I have already referred to the need to display CJK records at non-CJK terminals. There are over 600 terminals on our network, less than 10% of which will be capable of displaying CJK data. Yet all of these terminals can search the same database, and clearly a search on Library of Congress subject headings, or, for that matter, on the romanized form of an East Asian name, could retrieve a mixture of records with and without CJK characters in them. The relatively limited distribution of CJK terminals means that a reader of Chinese, Japanese, or Korean might have no alternative but to search the database -- through the likely intermediary of a reference librarian -- at a regular RLIN terminal. In addition, the non-reader of any of those languages who finds a CJK fish in his search net might well want access to the material anyway, to have translated or to use accessible contents such as tables, illustrations, etc.

As each RLIN terminal logs on to the system, a code will be associated with that terminal indicating its character-set capability. (An aside: RLG is investigating the feasibility of supporting other non-roman scripts as well as CJK, principally Arabic, Cyrillic, Greek, and Hebrew. The enhancements described here and in the paragraphs which follow, to RLIN software and to the MARC format, are almost all applicable to these other non-roman scripts. The only technique not generally applicable is the indexing/searching code normalization discussed below.) When a record in which there is a "character sets present" field is sent to a terminal, that terminal's capability indicator will be checked and the display tailored accordingly. For instance, if a user at a CJK terminal retrieves a CJK record (recognizable by the value in the "character sets present" field), then for most display formats the romanized version of a field which is also present in the vernacular will not be displayed; any field which occurs only in roman characters will, of course, be displayed. At the CJK terminal, the only time all fields--including equivalent vernacular and romanized versions -- are displayed is on the format used for cataloging, that is, a fully tagged and content-designated representation.

Conversely, the user at a non-CJK terminal who retrieves a CJK record will get all the romanized data displayed. Cataloging rules within RLG stipulate that, at the very least, there must be parallel entry of the title and statement of responsibility, edition statement, imprint, and series statements. Thus a basic bibliographic description is available to users at all RLIN terminals, whatever their (the users' AND the terminals') character-set capability.

Where indexable fields contain CJK data, those fields will be indexed and searchable in the vernacular. In fact, in order to take advantage of a particular feature of CCCII, some fields will be indexed twice, in normalized and unnormalized form. Within that portion of REACC which constitutes the Chinese character domain of CCCII, that is, not JIS codes with a prefix byte, or PRC codes with a prefix byte, etc., all recorded variant forms of a given character share the same values for bytes 2 and 3 of their code. Only byte one, the byte indicating the "plane" on which the character resides, differs, and it differs by fixed amounts. (Given the present audience, I am assuming a basic understanding of CCCII.) It is possible, therefore, by a fairly simple arithmetic manipulation, to convert any code within this domain to that of the traditional form of the character, our
"normalized" form. This then makes possible two different approaches to indexing and searching: if one knows precisely the characters in which an author's name is represented, then by searching the unnormalized index -- in which the data from the field being indexed is not transformed as described above -- one insures that only that form will be retrieved; but if the searcher isn't sure about the form, or, perhaps a more common occurrence, if one is searching on words in titles, then using the normalized index (and, incidentally, entering the search request using any form of the characters) will retrieve all instances of the name or word regardless of the way they were originally recorded. All this goes on only in indexing and searching: the characters in the actual bibliographic records -- the things stored and retrieved -- are never normalized.

Conclusion

Shortly after I wrote the foregoing, the first cluster controller and four RLG CJK terminals were delivered to us at Stanford, another concrete step in the realization of our goals in this project. Over the next few months programming, documentation, testing, and training will bring us ever closer to the day when the online cataloging of East Asian materials, and the integration of descriptions of those materials into a very large and invaluable bibliographic database, will--among the RLG membership at least -- be an everyday reality. At that point we will have achieved a less tangible but equally important goal: nothing less than -- as the President of RLG put it when she saw the first RLG CJK terminals -- "a revolution in East Asian librarianship in this country."