ISO/IEC 10646 and the Construction of the Circumpacific Documents Information Network

ZHU Yan

National Library of China

CURRENT SITUATION AND PROBLEMS

Viewing from the language Processing, Works that libraries & information departments with certain scale perform, is a processing of multilanguage. Because the books and documents they collected and presented to the readers are mutilingual. Even if documents published by native or local, foriegn characters and symbols still could be found besides native's own characters. Not only the text, also this is common on the title. Therefore, for libraries & information departments to descriptand report these documents, using of multilanguage becomes a must.

Before the adoption of ISO/IEC 10646 Universal Multiple-Octet Coded Charaters Set (UCS), computer systems currently in use are mainly bilingual systems, in fact, they are double character set transforming system. For instance, either processing English and Japanese, or processing English and Chinese, but the operating system do not support processing three of them altogether. Besides English (ISO 646) and several charater sets which are ISO standard, most of charater sets are still native or local standard. Furthermore, these standards are mainly used as information interchanging between systems. As for the inner character code system and processing machanism of the system, different computer manufacturer takes a different way. All these mentioned above give trouble to the on-line exchanging of inforamtion, make the development of liberaries & information departments and the establishing of multilingual information system become hard, and bring difficulty to the future construction of transnational and transregional circumpacific information network.

ISO/IEC 10646's BORN PROMOTED INFORMATION PROCESSING INTERNATIONAL LINGUALIZATION

In recent years, with the efforts of computer scientists, information processing specialists, and language experts of a global scale, ISO/IEC 10646 — a international unicoding charater set which compiling characters and symbols of all nations and regions — formally born at June, 1992, and issueing documents at March, 1993.

The written form of this charater set's applications on global language and symbols are: representation, transmission, interchanging, processing, storage, input and presentation. The characteristic of it are: multilingual, Multiple-application, mutiple-octet, C.J.K. ideographs character unicoding. It ends up the messful situation brought by different nation, different region, different company and different industry which all taken their own coding system, replaces it with standardized harmony of all-code-in-one; Moreover, it overcomes the troubles such as unequal in size, designating, invoking with extreme difficlty caused by the former character code, and makes the computer's software systems core independent from the charater. The new charater set leads to new datatype, brings about the innovation of operating system, promotes the born of international lingualized computer system. It is no wonder that the adoption of new charater set will accelerate the internationalization of software. All those excellent international softwares, including library information processing software package, no matter they are at systemlevel or application level, they all can move into the international lingualized environment without modification on the base level. The new charater set is both native and international. It is native because it compiles all standardized native charaters, and its font and style are native; It is international because when the user use native charaters, he can also use foriegn characters and symbols conviniently at the same time, and he canuse them together interchangeably. For example, C.J.K. ideographs character, is Chinese, Japanese, and Korean coexisting, simplified, complex, and variant forms coexisting, vocabulary collected reaches 20,902 ideographs charaters, whereas still can be expanded to fulfill further need.

In the international lingualized computer environment, construct circumpacific information network will be easy. In such case, libarian can use his native subject word heading documents from all nations, data of all kinds of language can be put into one same database, and managed by one software, readers even can use native subject word searching any nation's document. A transnational, transregionsl and mutilingual information system of real public, open, and network will provide the services.

WHAT SHOULD WE DO ?

Presently, some famous computer corporations are doing research on international lingualized brand new systems, making modifications and reformings on OS, programme language, DB, SQL, SPDL, SGML, and OSI, etc. But users who use the systems (especially library information system) are still needed to cooperate and discuss with the corporations on how to move the current software to new environment smoothly whereas least changes and least costs spent. This is particularly important to those library information networks which have been running for years, and have complete functions and still are effective. They should change the international lingualized environment as to let more user adopt this system and network, eliminate those low performance systems, therefore make information processing and information sharing step on a higher level.

Still we should study on the input method oriented 20,902 ideograghs characters. The new method do not only keep the quick & effective features current input methods have, but also can provide practical windows for inputting large sum of added nonfrequently using Chinese charaters. Current Japanese input method is based on the pronunciation of Japanese Kanji characters, this is effective to more than six thousands frequently using ideographs characters, but will be difficult facing 20,902 ideographs characters. It is considerable that one should stand on the basis of keeping current input method (i.e. facing frequently using Chinese charaters still based on pronuciation) to add more input methods based on shapes (e.g. strokes, radicals, etc.) as to find input method of nonfrequently using ideographs charaters. The reason is that although nonfrequently using ideographs charaters are large in sum, they are seldom used, enable to input but speed, will not affect the whole inputting efficiency. This advice is only for reference to Japanese experts. Also one should study on the input method native used to input foreign characters, for instance, Chinese charaters input method common Japanese could use, Japanese Kanji characters input method Chinese could use; etc.

For the deployment of abundant and practical ideographs character information processing supporting softwares to future's international lingualized computer system, I advise that China, Japan, Korea's experts will edit C.J.K. 20,902 ideographs characters' attribute dictionary together. The dictionary should include China, Japan, Korea's ideographs characters' pronunciation and ideographs characters' radicals, strokes, shapes and such necessary attribute informations.

Also we should preparing actively for the transferring from current information database to international lingualized system. Since ISO/IEC 10646 has give out relationship charts between the new code and the old code at its ideographs characters' chapters, data changing at future will be no longer a problem. But the relationship charts between non-Kanji-characters & symbols in JIS and the one in new standard have not been layed out, hence, we should completed this relationship charts. One more serious problem left is whether we should strive for unifying data structure in the circumpacific information networks whereby this chance of old system's reforming to international lingualized new system. For history's reason, all nations are not unified at their data structure, but observed thoroughly you could find, since each nation has been based on or closed to ISBD, basis of unifying data structures does exist. Unifying data structure will benefit realizing information's transmission and utilization between systems under circumpacific information network at OSI higher level's protocol. Nevertheless, how to unify, unify to where (for instance, is it considerable to unify them to UNIMARC?), still need the authority to discuss and to make decision on it.

Reference:

1 UNIMARC Manual/IFLA Universal Bibliographic Control and International MARC Programme.

2 ISO/IEC 10646-1: Information Technology-Universal Multiple-Octet Coded Character Set (UCS) 1993 (E).

3 ISO/IEC 10646 and Chinese DB/Zhu Yan (International Seminar on Chinese Document Database Feb. 27th, 1995 — Mar. 4th, 1995)

Requirements for the Internationalization of CJK Systems

Jack CAIN

ISM Library Information Services

Abstract:

This paper will discuss the following 7 topics:

1) a discussion of the implications for a CJK network of the recently developed protocols which provide a standardized basis for the searching and retrieval of bibliographic data. These protocols provide what are now termed "seamless links" between computers which may be completely incompatible in terms of hardware and software. The U.S. search and retrieve protocol is called "ANSI Z39.50".

2) in the communication of CJK data, the need for data identifiers to flag script, language and romanization systems so that such data can be correctly processed upon receipt. Such identifiers are not currently present in any CJK data communication.

3) the need for an international mechanism to establish and approve variant character linkages is examined in this paper. The need for international cooperation on this topic is stressed.

4) a special problem specific to the encoding of data in South Korea's national character set (KSC 5601) is discussed. Moving Chinese character (Han Zi) data into (and perhaps out of) this national standard requires a knowledge of the relationship between the encoding and the character's pronunciation in Korean. What mechanisms can be developed to cope with this international exchange of data?

5) the paper discusses the need for the internationalization of CJK data entry methods. Input methods for CJK data have been created in each CJK country but such methods are suited only to the requirements of that one country. How can data input methods be internationalized?

6) an examination of the possibilities and requirements for international agreement upon algorithms capable of supporting the automatic conversion between scripts — kana to hangul; Bopomofo to Pinyin, etc.

7) finally, a discussion of the limitations of the currently available CJK hardware and the software supported on this hardware is provided. What is now supported; what is required but is still missing.

Introduction

The purpose of this brief paper is to highlight a few of the specific areas of difficulty which will be encountered in the development of requirements for the international exchange of data in the Chinese, Japanese and Korean (CJK) languages. The resolution of these and other problems becomes daily more urgent as the amount of CJK data being created increases and the need to exchange this data internationally also increases. Databases of very significant size already exist in Japan and are rapidly being developed also in China, Taiwan and South Korea. As yet there has been very little international flow of this data outside of the borders of the countries in which it is being created. However, as part of the general trend of internationalization in the world today, it is inevitable that there will be an ever increasing demand to exchange this data between Asian countries. As soon as data begins to flow in this way, difficulties will present themselves in the manipulation of this data within environments quite foreign to its original environment. Addressing these difficulties is then the subject of this paper.

1. Search and Retrieve Protocols

Although the concept of linking disparate bibliographic systems running incompatible software and hardware has been seriously discussed for over 20 years, it is only in the past 4-5 years that real, practical development has taken place. The previous lack of progress was largely due to the lack of an agreed-upon standard and the recent developments have been made possible by the publication of ISO's "Search and Retrieve Protocol" and the corresponding U.S. standard known as "Z39.50". These protocols allow bibliographic systems to send and receive messages and search results in a neutral "language" which is not dependent on any one implementation of a bibliographic system. When using these protocols, the end-user may in fact be unaware that they are searching a remote system in a different city or a different country.

A pioneer in the U.S. in this area has been the local library system vendor "DRA" (Data Research Associates). Another example is the link recently announced between the U.S. bibliographic utility RLG (Research Library Group) and ISM Library Information Services (formerly Utlas International Canada). This recent development (implemented September 1994) is significant in the enormous size of the two databases which are connected in this way (22 million records in RLG; 50 million records in ISM). Although this connection has so far been used only to transfer data in Western languages, a CJK version of this link is now under development and a demonstration of the CJK link in test mode is being conducted later this week in Tokyo. So far as I know, this will be the first example of the successful transmission of CJK data via a Z39.50 protocol link in which the CJK data is fully displayable in both the source and target systems. Note that in the case of data transfer between RLG and ISM the internal character set handling on these two systems is completely incompatible; RLG uses 7 bit EACC code with escape sequences and ISM uses a private 8 bit character set with no escape sequences. In addition, RLG uses US MARC whereas ISM uses Japan MARC in its Japanese (Japan CATSS) system from which this link is being made. As with other Z39.50 linkages, the two systems are based on completely incompatible hardware (RLG has Amdahl; ISM has Tandem) and use completely incompatible application software.

The advent of search and retrieve protocols to link bibliographic systems may well be the most significant development since the creation of the MARC format. Certainly, it will be necessary to consider the implications of this development for all existing systems. Formerly, data was expected to be exchanged in batch mode by tape or file transfer. Now, data will be expected to be exchanged instantaneously online. This new mechanism will mean that problems which have been unsolved up to now, some of which are discussed in this paper, will become even more urgently in need of resolution.

The Z39.50 protocol is in fact expected to be expanded to provide for support in areas not now covered by the standard. One important area not addressed in the current standard is that there is no specification for announcing the character set being used in data transmission. Lacking this specification, the handling of character sets becomes entirely a matter to be agreed upon and handled by any two agencies which are exchanging data. Since a link can be made between any two systems using the standard protocol it is very important that provision be made to negotiate the character set being used in the data which flows between the two systems otherwise any data which is not in ASCII is likely to become garbage after transmission.

2. Script, Language and Romanization System Flags

In the transmission of CJK data, it will be necessary for a number of reasons to have one or more of these three indicators carried with the data. The reason for this requirement is to assist in the manipulation of the data once it has been transmitted to the target system.

For example, an enormous amount of development has been going on over the last 10 years in the development of computer fonts. This development has been taking place in the U.S. as well as in Asian countries. And in this development there are with certain characters significant differences between character shapes which are chosen for the fonts depending on whether is data is in Chinese, in Japanese or in Korean — the same character may have more stokes or may be drawn in a completely different style. Therefore when data is being transmitted, in order to display it in an acceptable font it might, in certain cases, be very important to know which language the data is in; this would be especially true in cases of files containing mixtures of data in more than one language. The presence of a flag for language of data would allow for acceptable font displays without human intervention.

An example of this case is the drawing of the Chinese character element meaning "grass"; in Taiwan, it is always printed with 4 strokes not three and in Japan it is always 3 strokes not 4.

Script and romanization flags may also become more and more important with the increasing amount of data transmission since it is common to have data in a given language transcribed into a different script. Such flagging would then facilitate the recognition of data and the possible machine manipulation of that data into a form more helpful for the user. Just one example of this is the manipulation performed by ISM when US MARC data is converted to Japan MARC format for use in Japan. In this case, the romanized forms of Japanese data are converted to provide katakana forms. This manipulation works reasonably well but falls down because the US MARC data does not indicate the language of the data on a field level but only on a record level. This means that when the conversion software encounters a record which is in the Japanese language but in which one field is not in Japanese but is instead in say Chinese, then the software which is attempting to provide katakana forms from the romanized forms will come up with an unpredictable result. To avoid this result we make some attempt to recognize that the data is not Japanese by having the program search for certain letter combinations which cannot exist in Japanese romanization — such as "tz", "hs", etc.

3. Variant Character Linkages

The universal nature of Chinese characters is such that those countries and languages which use them share what is sometimes called a "kanji culture". One of the practical advantages of this kanji culture is that a great deal of data may in fact be intelligible in its written form even when it is written in another language. However, the existence of variant characters in the modern standards for different languages (as well as within one language) can be a very frustrating barrier to communication. The ability to automatically convert all Chinese characters in a given file to the basic set for the language of the person who wants to examine the file would be indeed a useful feature. However, one of the barriers to such conversions is the lack of international and sometimes even national agreement on which characters could be converted to which characters. If such agreements could be reached then a file of Chinese data could be viewed by someone in Japan who does not speak Chinese using only joyo kanji forms. Such conversions would also be useful on demand within one language when one is viewing data which happens to use older variant forms. Of course such conversions can be made manually but the important thing is to have agreed upon tables which would permit such conversions to take place automatically and instantaneously.

A further practical advantage of such conversions is that they can permit, for example, Chinese data to be almost completely displayed on a device designed only to display Japanese — and similarly, for example, to display Japanese data on a device designed only to display Chinese or Korean.

Of course such conversions do not work for every character in the language but they do work for a great many very common characters which is why they would be extremely useful to the many people — librarians, scholars and information processing professionals — whose everyday tasks involve handling international data.

4. Moving Korean Hanja Data

A very difficult problem exists in the moving of Chinese character data used in files of Korean language data. This problem results from the fact that the Korean national encoding scheme KSC 5601 encodes the same graphic form of a Chinese character more than once when that character has more than one pronunciation in the Korean language. This is a very practical coding technique for sorting purposes in Korea itself but presents a real problem for moving Korean data which may have been encoded in other schemes (such as the U.S. coding scheme known as "EACC") into the Korean national scheme. Similarly, once data was moved out of the Korean national scheme and into some other scheme not based on Korean pronunciation it could not be moved back again into the Korean scheme. (Compatibility zones in Unicode and in ISO 10646 should allow conversion between these two standards and KSC 5601 but once the data is moved outside these standards the same problem exists.)

This difficulty in the Korean national scheme as it applies to the international movement of data is also a problem within Korea itself when it comes to the storing of data in Japanese and Chinese. In the case of data in these two languages, it is not at all sensible to store the data based on the Korean pronunciation of the Chinese characters which appear in data in these two languages.

In order to solve this problem of the transmission of data in Korean, very complex software would need to be developed. This software would need to be capable of consulting a large dictionary of Korean words and also be capable of performing linguistic analysis of grammatical contexts in order to determine the Korean pronunciation of characters which have more than one reading and based on that analysis assign the correct Korean encoding. It is unlikely that this software could work correctly in 100% of the cases. One further factor in the operation of such software is the problem mentioned above of encountering data in say Japanese or Chinese within the Korean file — for example, the translation of a Japanese work into Korean where the author's name is Japanese, not Korean.

5. CJK Data Input Methods

Much work has now been done on the creation of fast, efficient and user-friendly input methods for data in Chinese, in Japanese and in Korean. However, up to now, this development has been localized within specific markets and has not been internationalized. What I mean is that good methods exist for inputting Japanese in Japan and Chinese in China and Korean in Korea. But no methods exist for inputting Chinese in Japan or Japanese in China. Such combinations present their own very real problems which are quite different from the simpler cases of one language in one country. For example, when inputting Chinese data in Japan it would be very helpful to allow the Japanese inputter to type the Chinese characters using Japanese pronunciation and only in those cases of characters which do not exist in Japanese would it be necessary to switch to a Chinese (or Roman) input method. Similarly in China, as well as in Korea, allowing the input of Japanese characters as they are used in Japan but with Chinese, or Korean, pronunciation would be a very helpful product; at present such products seem not to exist.

6. Automatic Conversion between Scripts

In section 2, I spoke of the need for data encoding by language, script and romanization scheme. However, this requirement in the distribution and exchange of data is only part of the difficulty in the internationalization of data flow. Romanization is only one of the many possible transliterations of scripts but it is one for which standards tables have been developed by ISO and other bodies for a great many languages and scripts. However, many other possible and useful conversions do not have (to the best of my knowledge) internationally recognized tables — for example,

Hangul to kana and kana to hangul
Cyrillic to kana and kana to Cyrillic
Arabic to kana and kana to Arabic
Bopomofo to kana and kana to Bopomofo (Taiwan)
Hangul to Bopomofo
Hangul to Cyrillic
etc.

In order to allow for the machine manipulation of these scripts into a writing system best known to the user of the data, it is necessary to establish agreed-upon standards for this process. Once standards have been agreed upon then it is a relatively simple matter for computers to provide for on-demand presentation of data in a script which the user requests. Without agreed-upon standards however, the progress in the practical application of such conversions is greatly impeded.

7. Current CJK Vendor Support

As a conclusion to this brief paper, I would like to review the current situation with environments which support CJK processing. By this I mean the availability of support for the simultaneous processing of all three of these languages. In spite of major advancements in the processing of each of these languages separately there is still a minimal amount of commercially available support of multilingual environments. For example, there is now support in Japan, on the level of word processing equipment, for Chinese or Korean but, so far as I can determine, there are not any products in Japan which support the processing of all three languages simultaneously. And in the case of Chinese, the products which are available in Japan do not allow the simultaneous recording of both traditional and simplified Chinese characters in the same file with separate encodings. The need for support of both traditional and simplified Chinese becomes more urgent as publishing practices change and we begin to see traditional character publications from Beijing. I will mention 3 products from outside Japan which are of interest. Perhaps there are other commercially available products but these are the only three of which I am aware.

7a: Unicode and ISO 10646

The Unicode (1992) and ISO 10646 (May 93) standards have now been published for about 2 years. These standards are the only character set standards which support a truly international multilingual environment for the exchange of all types of data. When these standards were first published it was expected that products based on them would be forthcoming very quickly. However, this has not been the case. At the Unicode Implementer's Meeting in Santa Clara, September 8-9, 1994, it was clear however from reports by companies such as IBM, Microsoft, Adobe and others that a great deal of work is actively being done to prepare Unicode-based products for the market. Windows NT for example claims to already supports Chinese and Japanese but I have yet to have this verified by someone who has actually used the product so I am unaware of what level of support is actually provided inside Windows NT for the input and display of Chinese or Japanese data.

7b: MASS

MASS is a product from the Institute of System Science at the National University of Singapore and has been developed to run on UNIX workstations such as Sun (OS 4.1.X) and Solaris (2.X). This product has been chosen by Australia in their national CJK project coordinated by the Australian National Library. This product is Unicode-based and provides a true multilingual environment. The following languages are currently supported: Arabic, Chinese (Simplified &/ or Traditional), French, German, Greek, Italian, Jawi, Korean, Spanish, Tamil, & Thai. One of the most interesting features of this system is that it allows users to switch input methods independently of the language of the data being entered as well as independently of the language of the interface being used. (The language of the interface is also switchable). This is the first system I have seen that allows the input of Chinese data using Japanese readings of the Chinese characters. The product supports about 22, 000 Chinese characters — that is, all characters present in the Unicode standard; this therefore includes all characters in JIS and in the basic standards from the People's Republic of China (GB2312-1980) and from Taiwan (Big 5 or CNS 11643).

7c: JOIN & Pacific Data

From Taiwan there are two products based on the CCCII character set standard which offer support for the simultaneous use of Chinese, Japanese and Korean. The names of the two companies offering this support are "JOIN (Yong Chi)" and "Pacific Data (Chang Tai)" both with offices in Taipei. At present, it is not possible to use these products for Cyrillic or for European languages but the National Library in Taipei is now spearheading a movement to upgrade the CCCII character set with new codes and input sequences which are at present lacking for these languages; this development is expected to be implemented by the Taipei vendors in 1995. The CCCII character set is much larger than that supported by Unicode and now includes about 70, 000 codes although some part of this large number are duplicate graphics intended to show variant character relationships. Input sequences for Chinese characters are available only by Chinese input methods thus the input of Japanese data which is not kana is very difficult for Japanese speakers who do not also speak Chinese fluently and with accurate pronunciation.