APPLICATION AND DESIGN CONSIDERATIONS CJK FOR INFORMATION INTERCHANGE CODE

*Chen-Chau Yang**

Abstract

The most urgent need of the information interchange code for facilitating the characters in Chinese-Japanese-Korean material appears to be in library applications. One of the problems in this area is mainly caused by the Chinese characters (KANJI). Although many of the characters possess the same semantic meaning, they have different shape and pronounciation; or some characters have the same shape but different semantic meaning and pronounciation. In this paper, we will first discuss the application requirements of CJK information interchange code in library environments. Then we will try to identify the important problems that must be considered in the design of an information interchange code for CJK applications in library environments. Finally, we will review the design of the Chinese Character Code for Information Interchange (CCCII) and the feasibility of adopting it as an international standard for CJK information interchange code.

Introduction

The exchange of information between two computers can be carried out either on-line via telecommunication channels, such as cable or satellite, or off-line through storage media, such

*Chen-Chau Yang, Prof., Dept. of Electronic Eng. & Tech., National Taiwan Institute of Technology, ROC.

as magnetic tape or floppy diskette. No matter whether it is on-line or off-line communication, there are a number of steps and certain data formats to be followed between the two communication computers, this is the so-called "protocol". The procedures of communication, e.g. establishing the connection, the sending of data, acknowledging the receiving of data, and terminating of the communication, etc., are well defined in telecommunication standards. Some of the data formats are well defined too, e.g. the MARC format. An information interchange code is also one of the data formats, which is used for coding the characters. For example, ASCII code is the standard for coding English characters. Information interchange codes should not be confused with the code used inside the computer system, although they may be the same. Many computers use a special internal code because of their hardware architecture, e.g. 6-bit character code is used in CDC computers as their internal code. However, the conversion from internal code to interchange code, and vice versa, usually can be done by the computer automatically under user-control.

Theoretically, we can use any code for exchanging data between different computers. If the communicating computers all agreed to use a certain internal code for exchanging information among them, then there is no need to devise a separate information interchange code for these computers. However, generally, different computers may use different internal codes, and hence for at least some of these computers the code conversion has to be done inside the system. In addition, it is highly possible that certain binary bit patterns of the chosen internal code may be used as control codes in some of the computer systems, this will require a lot of processing time to verify, as every byte of data must be checked against the list of control codes. Therefore, technically we have to use a separate information interchange code for exchanging data between different computer systems. In this case, although every computer involved in the communication will have some overheads on the code

conversion task, it will largely simplify the design of the application software. The overheads on code conversion are very small, because it can be done by hardware and processed in parallel with the application software.

In the case of KANJI (Chinese characters), As far as information interchange code is concerned there is no international standard. The problem is complicated further by the fact that there are so many different internal codes used for representing KANJI inside the computer, almost every different KANJI input method will result in different internal code. Hence, it is necessary and critical to devise an internationally acceptable standard in this area if we are serious about exchanging bibliographic records among libraries by either MARC tape or through on-line communication.

Application Requirements in CJK Library Environments

There are a number of problems that must be considered when applying computers in the CJK library environment. Firstly, the character set is larger by several orders of magnitude than the most extended Roman or other alphabetic character set. Secondly, the character set must be perpetually extensible to cover those characters that are either created or invoked by a user, as these characters may be unknown to other users. In other words, the character set is a growing set, and eventually will cover all the characters which have ever appeared in written material when full text processing is economically feasible and needed in practice. Thirdly, the character set must be capable of being fixed at any single point in time but still allow the addition of a certain number of character codes that are not included in the set but are needed for the processing of some bibliographic records. There must be procedures to add to the set these characters which have been temporarily assigned codes. Finally it is desirable that there is a method for linking characters that

are different in shape but have the same semantic meaning, i.e. variant form characters. This requirement results mainly from cataloging rules: whatever appears on the title page of a book would have to be reflected in the corresponding bibliographic records in the computer. In other words, we need to maintain the original form of the title page of a book but still be able to display the title page in characters which are familiar to local users. Therefore, when designing an information interchange code, the capability to link variant forms of characters must be included in the code. The actual transformation from character to character should be done automatically using software this is dependent upon on the specific needs of each independent library environment. The questions of what should be considered as the variant forms of a character and how many variant form characters should be included in the character set are linguistic problems and will not be discussed further in this paper.

CJK Coding Scheme Design Considerations

In the design of a CJK information interchange code, the first problem that must be considered is the international acceptability of the code. It should be acceptable by various brands of computers and compatible with existing international communication standards. Because we cannot expect that all the computers that will use the code can handle high level communication protocols, such as the X.25 of CCITT or SNA of IBM, in which data can be represented in binary bit sequence without considering possible confliction with communication control codes, we have to follow the lowest level standards, namely, the standards of ISO 646 7-bit code and ISO 2022 multiple bytes extension. The details of these two standards can be found in related ISO publications and will not be repeated in this paper.

The second problem that must be considered is that the coding space should be large enough to cover all the KANJI

characters that ever appeared in written material and still have space to include characters that may eventually be created by any user (legitimately or illegitimately). Roughly, there exist more than fifty thousand KANJI characters with an unknown number of variant forms; according to the estimation of the Chinese Character Analysis Group there are about thirty thousand variant form characters in addition to the fifty thousand traditional characters.

Finally, the information interchange code should fulfill the application requirements as mentioned in the previous section.

The abovementioned points have been considered in the design of the Chinese Character Code for Information Interchange (CCCII). We shall briefly review the major characteristics of CCCII in the sequel. CCCII was designed to be fully compatible with the ISO 646 standard. The escape sequence for CCCII is proposed to be ESC, 2/4, 4/2, this is a method for extending graphic symbols with multiple bytes representation. In CCCII, each graphic character, i.e. Chinese character, is represented by a 3-byte code. With the 3-byte coding scheme, the coding space can accommodate more than eight hundred thousand graphic characters. Fig. 1 illustrates the three dimensional structure of CCCII. The first 15 sections in plane 1 are reserved for special purposes, their usages are shown in Fig. 2. When implementing the coding structure, the planes are reorganized into 16 layers, with 6 planes in each layer except for the last layer which contains only 4 planes. Fig. 3 gives an illustration of this coding space organization. The reason for adopting such an organization is the requirement of being able to link variant form characters to their corresponding normal form characters, as discussed in the section of application requirements in this paper. The variant form characters are placed from plane 2 up to plane 15, depending on how many variant form characters exist for a normal form character. By this arrangement, in the character code, the right two bytes of a code are the same for a normal form character and for its corresponding variant form characters.

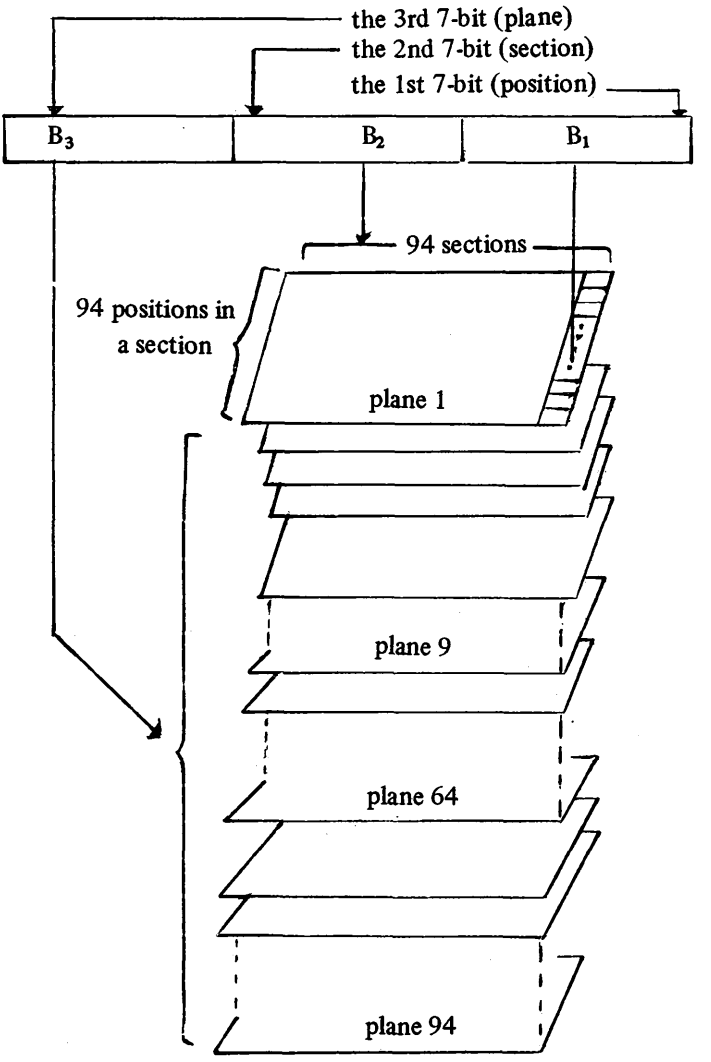

Figure 1. The 3-dimensional structure of the three 7-bit bytes coding space.

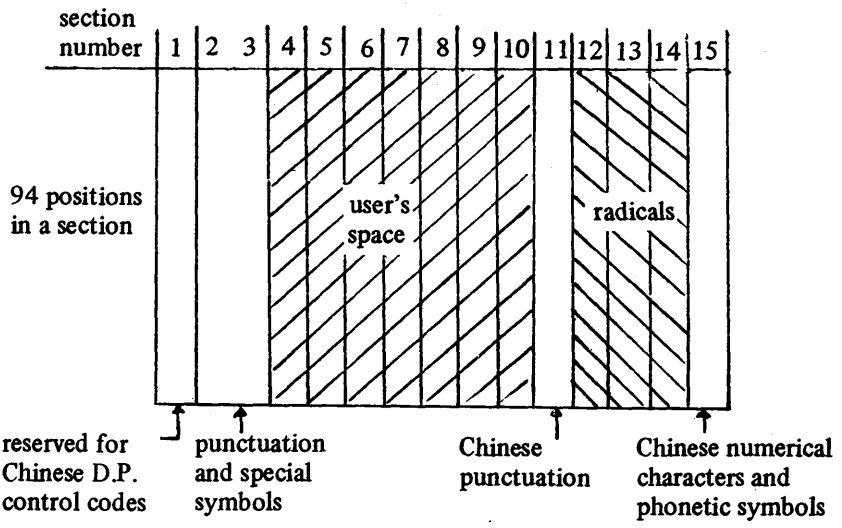

Figure 2. The structure of the first 15 sections in plane 1.

Hence, we can link a variant character to its corresponding normal form character simply by using the two rightmost bytes.

Currently, CCCII contains more than thirty thousand Chinese characters, one third of which are variant form characters. These characters are collected from the major data processing centers in Taiwan, libraries, and those published as the frequently used characters by the Ministry of Education. The Chinese Character Analysis Group is still collecting other characters that will eventually included in CCCII. From the above discussion, it can be seen that the design of CCCII actually considered every important application requirement and design consideration.

Since the publication of CCCII in April 1980, the Research Library Group has adopted the coding structure of CCCII as the basis for developing their East Asian Character Code in the Research Library Information Network. The CJK project of RLIN is the first implementation of the CCCII code structure

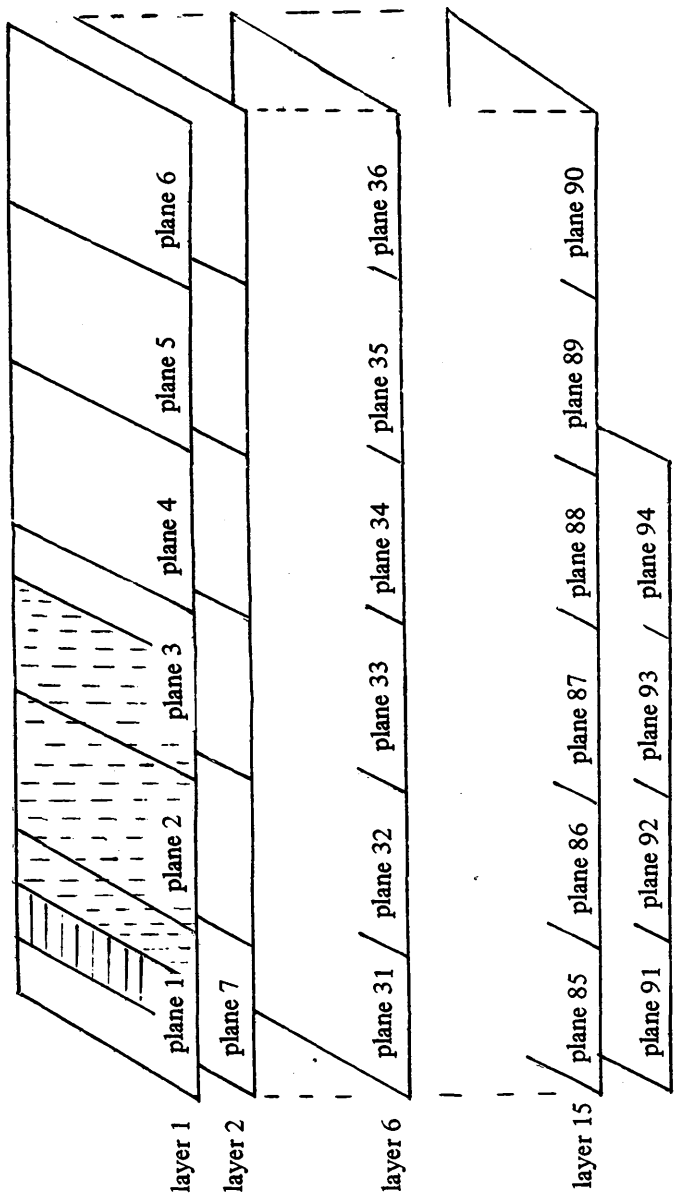

Figure 3. The Structure of CCCII.

in an operational environment. The details of this project and the RLIN CJK system can be found in their technical report and will not be discussed further in this paper. The point is that the RLIN CJK project has successfully demonstrated the feasibility of adopting the CCCII structure as a coding standard for CJK vernacular material in an automated bibliographic system.

Concluding Remarks

In this paper, we have discussed the application requirements and the points to be considered when designing a CJK information interchange code. There are some other problems that must be considered when implementing the interchange code. For example, the characters in the set may have to be arranged according to certain order, such as the radical and stroke count of the character or the pronounciation of the character, or the characters may have to be grouped by their usage frequencies. All these problems are application dependent, and must be solved from an applications point of view. As far as the coding structure is concerned, these processing requirements pose no difficulty at all. For example, if the characters are arranged according to radical first then stroke count, and we need to sort the material strictly according to the stroke count of characters, the simplest way is to devise a conversion table containing the interchange code and the corresponding internal code of each character, in which the internal codesare ordered by stroke count. By this approach, any internal processing of the characters can be handled with some overheads on code conversion. However, the important point is that the information interchange code is standardized and the code conversion task is straight-forward. From this point of view, CCCII code structure provides an ideal underlying structure for the development of the CJK information interchange code.