PERSONAL NAMES AND THE CHINESE CHARACTER CODE FOR INFORMATION INTERCHANGE, VOL. 1: (CCCII/1) – ADEQUACY AND IMPLICATIONS†

*James E. Agenbroad**

Introduction

This paper has two purposes. First it reports the results of a little experiment to test whether or not the first volume of the *Chinese Character Code for Information Interchange* (referred to hereafter as CCCII/1) contains the characters needed to represent a small sample of Chinese personal names. The second purpose is to consider some of the available alternative responses when the characters needed for a particular purpose, e.g., to catalog a book, are absent.

The Test

There are two elements to my experiment, CCCII and the sample of personal names. I will describe each briefly.

Others here could and I hope will give a better and fuller explanation of CCCII. To me, CCCII is a very successful attempt to devise a scheme of internationally acceptable computer codes and standardize the assignment of these codes to Chinese

---

† Paper presented at the International Workshop on Chinese Library Automation, Feb. 14-19, 1981.

* James E. Agenbroad is Computer System Analyst, Automated System Office, Library of Congress.

characters. Only after consulting it in detail for this paper did I begin to appreciate how much effort its preparation must have required. The more I used it, the more impressed I became. Written Chinese requires thousands of different characters and, consequently, thousands of different codes. Traditional computer applications with numeric and alphabetic data have needed less than a hundred characters and less than a hundred different codes. To provide for the greatly increased number of codes, three of the standard codes are used to represent each different character. Using combinations made from any three of 94 different codes makes it possible to define 94 or over 830,000 different codes, which should be enough even for Chinese. While volume one of CCCII established a scheme for 830,000 different codes, it assigns codes to only 4,807 frequently used characters. As this is written I have not seen volume two, but I understand it will be available at this workshop. If it continues the excellence of volume one, it will be worth the wait.

While CCCII was established for bibliographic data the characters defined in volume one were frequently used ones. It occurred to me that it would be useful to see how adequate these characters would be for a sample of personal names both ancient and modern. Fortunately, when I was in Taiwan last June I was given a copy of *Portraits of the Greats of China.* It is published by the Dr. Sun Yat-sen Memorial Hall and contains brief biographies of one hundred Chinese. They were chosen as exemplars of traditional national spirit and extend from Huang Ti, a legendary ruler who lived 4,600 years ago, to Kao Chih-hang, a fighter pilot who died in 1937. It includes men and women, sages and statesmen. Each name as given on the Chinese contents pages was searched in CCCII to see if the characters needed to represent it were present. The hundred names used 225 characters; all but six were present in CCCII.

In the following list of names containing characters not in CCCII, the absent character has been underlined in the romanized portion.

1. *T'i* Ying, ca. 100 B.C. 緹縈
2. Mi *Fei*, 1057-1107 米芾
3. Chu *Ta*, 1626-1705 朱耷
4. Shen *Pao*-chen, 1820-1879 沈葆楨
5. Ch'iu *Chin*, 1875-1907 秋瑾
6. Ts'ai *O*, 1882-1916 蔡鍔

In one other case, Tsu Ch'ung-chi, 429-500, the second character of his name appears in the contents note written with the ice radical while in CCCII it appears in a variant form with the water radical. I will leave to others the question of whether or not 沖 is close enough to 冲 to meet the needs of catalog users. It is not a trivial issue. Since CCCII is arranged by radical, the character would be found only by looking under the water radical at row 5, column 7 of section 38.

Especially with older names the popular form of name was used on the contents page and searched by me, e.g., Lao Tzu (老子) and not a more official form of name which might be used in heading such as Li Erh (李耳). Since the popular form would still often be needed when recording the information on the title page, this should not matter. Contrary to my expectation the characters needed were more often unavailable for recent names than for older, i.e., pre-1500, names.

This brief survey of Chinese character availability highlights the great difficulty posed by applying computer techniques to Chinese. No matter how large the repertoire of available characters is, the need for additional characters will sometimes arise. A system designer may feel like the child who received a failing grade for spelling ten words wrong in an essay and said plaintively, "But look at all the words I spelled right."

The frequency with which the need to deal with absent characters arises will depend on at least three factors. First, the size of the character repertoire – the larger it is, the fewer occasions it will be found wanting. Second, the functions to be performed with the data – a character repertoire adequate for

preparing ephemeral records such as telephone bills, book orders and library circulation statistics may not be large enough for compiling a national bibliography, a dictionary or a concordance to a poet's works. Such works of more permanent value may require additional characters to make needed graphic, phonetic, sorting and semantic distinctions. Third, the source of the data – any specialized application, be it medical, chemical, historical or artistic will use a special vocabulary and require more different Chinese characters than a less specific application.

Absent Character Alternatives – I: Do Without It

When a needed character is not available in a system's character repertoire, the possible solutions fall into two broad categories – first, do without it, and second, add the character.

The following are some procedures falling into the "do without it" category.

1. Key an X or some other distinctive symbol in lieu of the unavailable character. Then paste or write in the needed character on the output document before sending it to the printer. This has several disadvantages. It is of no use for an on-line application – you cannot write on the terminal screen. Second, the keyed data may not be manipulable by the computer as intended, e.g., it may be impossible to sort it correctly. Last, you must maintain special manual procedures to keep track of which character you want to add where before printing. Errors will inevitably occur in a prominent portion of the publication.

2. Key an X or some other distinctive symbol and add a note referring to a reference tool in which the absent character can be found; e.g., "Substitute Mathews 1754 for X." Besides preventing accurate computer processing,

this procedure forces the reader to consult a reference book (which may not be readily available) before understanding the data.

3. Key an approximation of the needed character. This approximation could be:

   a. Phonetic – a character with the same sound and tone.

   b. Graphic – one or more characters which resemble the image of the absent character. For example one could key "ice" (氵) and "center" (中) in lieu of 冲 . (At the Library of Congress we have adopted a similar practice for the keying of letters used in certain African languages which are not on our Latin alphabet character set.)

   c. Semantic – one could key a character with the same meaning though it looked and sounded different; in effect translate the absent character into an available equivalent.

All these approximations are fraught with misunderstanding, because it may be difficult or impossible to tell the reader which type of approximation was used. If you key a phonetic equivalent which is read as a semantic equivalent the results may be ambiguous, misleading, humorous or worse.

Despite their disadvantages the above procedures or others may be adequate when only temporary data and less than total accuracy are required. The virtue of such solutions is that they are relatively easy for those who create data to accomplish quickly when a need arises. No changes to existing software or hardware are necessitated. Instead the extra burden is passed on to the final user who may suffer delay, confusion or disappointment – unless he understands the intended message, the system has failed. The frequency and enormity of such failures would be very hard to measure. Questions about the value of information and the

expense incurred by its absence are too large to treat here.

Absent Character Alternatives – II: Add It

We now come to the second category of solutions to the need for a Chinese character not available in CCCII; namely, add the character to the character set. This solution is philosophically preferable because it will involve no extra effort by the reader whom we seek to serve. Instead, it involves extra effort by those who develop and maintain CCCII and those who develop hardware and software which use CCCII.

There are three aspects to this solution:

1. Deciding to add the character.
2. Changing the software and hardware.
3. Notifying interested parties of the addition.

The purpose of CCCII is standardization of assignment of characters to codes to facilitate communication between computers. Hence the more people who agree on such assignments, the wider the communication. In CCCII, volume one, there are 658 codes set aside for user assignment. I do not know whether or not later volumes will provide more codes for users. Before any characters are added it would be very desirable to define the criteria or guidelines for when additional characters will be added. For example, are variants such as 冲 and 沖 both needed? As another example, presently CCCII has 48 characters containing the horse radical; must separate codes also be assigned to the same 48 characters which use the simplified version of the horse? The guidelines for local (i.e., user) space and true CCCII code assignments may not be identical. It would be preferable that they were, but the user-assigned codes might need to make fewer distinctions if the user's application were less sophisticated. For example, the distinction between 冲 and 沖 is more important to those who wish to sort by radical or stroke than it is

to those who will sort only by the romanized form of a character.

When characters are assigned codes locally from codes in the user space area, communication using these codes will be limited to those who are aware of and agree with the local character assignments. On the other hand, readers of books with title pages containing unavailable characters will be unhappy to wait for a desired book until the next volume of CCCII appears with the code definition of the needed character.

The conflict here is between fast local response to a specific need and a slower, standardized response which will facilitate the widest and easiest sharing of machine-readable Chinese character data. Both are desirable goals.

It is not enough to say this character which is new to the system will have this code. Once the decision has been made to add a character and a particular code assigned to it (characters in CCCII are in radical and stroke order), the character must also be included in any hardware and software system which uses CCCII. In order to do this it will be necessary to specify at least the following:

1. How the character will be keyed.
2. How it will appear on the screen – all terminals will need to have the instructions for "generating" the new character when its code arrives.
3. How the character will appear on paper – printing devices will need to be adjusted to respond to arrival of the new code. If sorting or retrieval will also be done based on Chinese characters, additional information about the character will be needed. This type of effort will require staff with both linguistic and data processing expertise. The design of a particular system will determine how difficult making the needed changes will be; but, nevertheless, the changes will be needed unless compromises of the kind previously described are acceptable.

Lastly, there is the matter of notifying all who need to know of new characters and assigned codes. There are two aspects of this. If the code is in the user area, only those with access to that system will need to know. If the code is in the regular part of CCCII, it must be more widely distributed. There are a limited number of codes in the user area. If several groups are assigning codes locally, the possibility for assignment of different characters to a single code is great, especially if the groups are on opposite sides of the world. This would mean that data from one such group could not always be correctly understood at another institution which had assigned a different character to the same code.

In addition to the time needed to achieve agreement on new characters and their codes this type of publishing tends to be slow and expensive due to both the relatively low demand and the high cost of typesetting. The cost could be reduced were an automated system for control of CCCII implemented. This could also reduce the mechanical aspects of slowness of publication. It would not help with the political and intellectual aspects of obtaining international agreement on the sensitive and complex issues involved.

As an attempt to alleviate the dilemma of fast but local assignment of codes for new characters versus slower but more widely accepted single source for doing so, the following procedure is proposed.

1. The user codes should be considered as temporary character code assignments pending publication of a final assignment in CCCII.
2. Portions of the user codes of CCCII should be allocated to major users to prevent conflicting assignments. One possibility would be to reserve blocks of ninety-four characters per major user goup.
3. All local code assignments would be forwarded annually to some central agency which would prepare new lists of standard CCCII code assignment. The location,

makeup and financing of this central agency is inevitably a politically sensitive issue – people feel strongly about their language. The originators of CCCII should definitely participate in such an agency. Whether or not it might be located elsewhere (e.g., Hong Kong or Macao) is a question I leave to others.

4. Local agencies would then update all occurrences of locally assigned older codes with the approved standard codes. The records containing the local codes might be marked so they could be easily retrieved. The local hardware would also need to be adjusted.
5. The old local codes would then be free for local assignment again.

Note that this is only an attempt to describe one possible solution to a difficult issue; other procedures might be more effective politically or technically or both. Three things are certain:

1. Agreement on a set of computer codes for Chinese characters is needed to improve bibliographic control of East Asian materials.
2. The set will need to grow.
3. Independent additions will breed confusion unless we begin now to establish orderly procedures for this growth.