A Missing Characters Description Language for Han Characters

Ching-Chun Hsieh
Institute of Information Science , Academia Sinica

Lin-Lin Wu
Department of Information Management , National Taiwan University

Ya-Min Chou
Department of Information Management, National Taiwan University

Abstract-*The purpose of this paper is to solve the interchange problem of missing characters. For decades, we always have to face the missing characters problem while using computers to process Han characters. Missing characters causes the information retrieval and interchange problem which is more serious in digitalize ancient books and Buddhist Sutra, however, there are only few studies on this problem. In this paper, we propose an interchange framework and Missing Characters Description Language(MCDL) - a language and protocol for describing the knowledge about Han characters to solve missing characters problem. An MCDL-based implementation is provided as well. Experimental implementation shows the missing characters problem can be solved successfully.*

Keywords: Missing Characters, Metadata, Document Interchange

1. Introduction

The missing characters problem occurs when one cannot find the intended characters in computer's character set. The problem becomes potentially overwhelming in the context of a logographic writing system. The major reason is current encoding systems are designed based on alphabetic and syllabic writing systems. For syllabic and alphabetic writing systems, a character is a writing unit and element of phoneme or syllable [1]. Because the number of phoneme is finite, we only need finite phonetic symbols to represent sounds of words for a language [2]. With these properties, most encoding systems assume the characters of all language are a closed finite set, however, this assumption is not correct at all to logographic writing system.

In a logographic writing system, each character is not only a writing unit but also an idea or concept [1], a typical logographic system is Han characters. Because the Han characters represent ideas, it will need many different characters to express lots of ideas. The complete Han writing system is expected to consists of 40,000-70,000 characters (accurate estimates are difficult) each representing one or more different concepts. This implies a requirement for more characters and a larger character set than those by alphabetic and syllabic writing systems. Because we need lots of characters, it is more likely to encounter missing characters if the character set of encoding system is too small.

Although a larger character set can be employed to reduce the occurrences of missing characters, we can not avoid it happened at all. As Wittern and App say[3]:"In East Asia, the problem of missing character is ubiquitous…It is clear from our work on electronic Chinese Buddhist texts that even Unicode will not significantly reduce this problem". Therefore, we can not solve the missing characters problem simply by using a large character set.

The previous studies on missing characters problem can be divided into two categories. The first category focus on how to represent missing characters. Regarding this problem, encoding systems including Unicode [14], CNS 11643[16], CCCII [17], GB18030 [15] reserve a private use area for assigning codes to self-created characters. However, independent users might assign the same code to different characters or different code to the same character, so it is difficult to exchange the documents contains missing characters. Hsieh [5][6] and Wittern[3][4] proposed two different regular expression of glyph to represent missing character. The benefit of this approach is it can be independent to encoding system, yet, applications use this method may be limited in terms of efficiency and flexibility.

The second category focus on how to interchange the documents contain missing characters and we also concentrate on this problem. So far, there are three approaches proposed in the literature: (1) creating a larger character set for interchange (2) dynamically generating characters (3)disallowing new characters to be arbitrarily added in the private use area. The first approach is adopted by many encoding system including Unicode [14], CNS11643 [16], CCCII [17]. However, the characters not in character set are still unable to be transmitted. The second approach only focus on displaying problem, but, editing and retrieving of missing characters are not considered [8][12]. The third approach is adopted by Quanzi database and Japan Mojikyo fonts database [7] which require applications for adding new character to avoid inconsistent allocation of an

identical character to character. However, an August 2003 survey states that only 13% of users will send relative information to apply adding new characters, whereas 48% of users still like to add characters in private use area[18]. In summary, previous works could not solve the interchange problem of missing character very well and still need be further improved and studied.

2. Definitions

In order to solve the interchange problem of missing character, there are some terminologies should be defined first, these terminologies include character, glyph and font.[5]

(1) Character: A character is a writing unit and abstract concept.

(2) Glyph:A glyph is the concrete structure of character. Each character may have many glyphs. The major difference among glyph is their structure. For example: the concept of clear has two glyphs(晰晢), and the concept of wisdom has five glyphs(智知恕曎朁)

(3) Font:A font is a group of rules that applied on glyphs. Each glyph has many fonts but the structure of different fonts for a glyph is the same. For example: Both 智(kaishu font) and 智(shiming font) have the same structure.

Sometimes, several characters might be represented by the same glyph. According to Chinese etymology, different glyphs are often used to represent near-homophonous characters that are unrelated in meaning[19]. The major benefit is it can significantly reduce the number of glyphs needed to express the concepts.

In fact, some missing character problems are missing glyphs not characters. For examples, we need a glyph 邋 (Bronzi script) which is not in Big-5 and Unicode character set. But, both 邋 and 歸 are glyphs for the same concept.

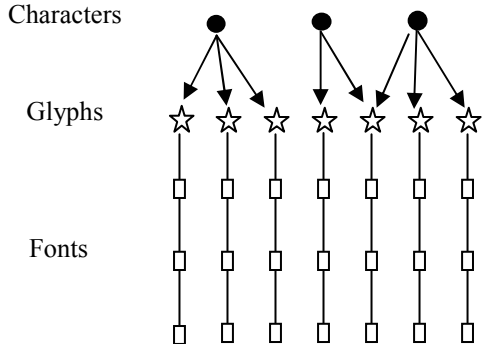

Figure1. Characters, Glyphs and Fonts[5]

3. Methodology

Why the missing characters interchange problem not be solved? The essence of the problem is that both senders and receivers don't have enough knowledge to processing missing characters, they exchange missing characters only relies on codes. If both senders and receivers have the knowledge of missing characters, they might be able to process missing characters properly. However, what knowledge about missing characters do sender and receiver need and how to represent the knowledge about missing characters?

We think that the most important knowledge is the structure of glyphs, because the main difference between any glyphs is their structure. In addition, the method of describing a glyph should be readable and identifiable by human and computers to make the process of missing character easily and flexibly. In this paper, we use the glyph expression developed by document processing lab of Academia Sinica[5][6] to describe the structure. The primary concept of glyph expression is that each glyph can be decomposed into several components or roots. These components or roots can use three operators include horizontal(⿲), vertical(⿱) and contain(⿴) assembled as any glyph. This means that glyph expression has generating ability to represent infinite glyphs. According previous study, there is only one pair of glyphs have the same glyph expression(畾=品⿱亖山,畾=品⿱亖山) in the Hanyu Da Zidian which contains 5,4000 glyphs, actually, they are the same glyph. Therefore, the glyph expression can be used as an identifier for each glyph, because there are no different glyphs have the same glyph expression.

We also need the knowledge about the relation among variants for the same character. These knowledge including time dependency, pronunciation of ancient/middle-ancient/modern, semantics, part of speech, all are extracted from Chinese etymology.

After these analyses, we develop a language MCDL to describe the knowledge for missing characters discussed above. The specification of MCDL is written using the syntax of XML and has the several features:

(1)It is easy to use in combination with other metadata markup languages.

(2)It is capable of representing the knowledge of missing characters.

(3)It is easy to be created, read, and modified.

Actually, the description of missing character by using MCDL is also a metadata and can be integrated with other metadata to meet the different requirements of applications. The relation of MCDL to other metadata languages is shown in figure 2.

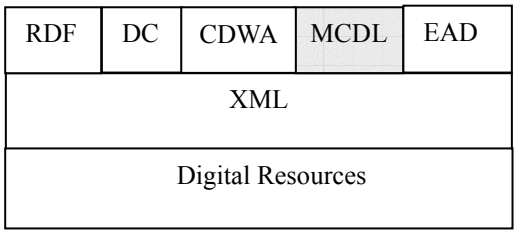

Figure 2. The relation between MCDL and metadata description language

In MCDL, there are many elements to represent the knowledge of missing character.

(1)MCDL element: The first element is MCDL element that is the root element for all elements and must contain at least one Character element. The MCDL element's attribute include the namespaces used by MCDL.

(2)Character element: Each missing character is described by a Character element which contains all the elements to describe missing character. These elements are Apply-glyph, Glyph-expression, Code, Font, and Variants.

(3)Apply-glyph element: The glyph selected to express a character will use Apply-glyph element to describe its glyph expression, code, character set and font information.

(4)Variants element: Each variant of character is described by a Variants element.

(5)Glyph-expression element: Both Apply-glyph and Variants have the Glyph-expression element described the structure expression of glyph for missing character and variants.

(6)Code element: Because using private use area to add new character is a most important method to represent missing character, we use Code element to represent the character code assigned to missing character.

(7)Font element: The Font element is a external reference to indicate where the font is.

When the character set does not have the same glyph of missing character, it is not necessary to create new glyph if there are variants with the same meaning. The first advantage of using variants is that users do not need to use private use area which only has limit space. The second advantage is it can reduce the interchange and retrieval problems cause by missing character. But the variants and the glyph of missing character might have different appearances that have to specify in order to distinguish them. These differences must be represented in Different element which is a sub-element of Apply-glyph and Variants.

The differences of variants and the glyph of missing characters are classified into four categories:

(1) Stroke:The variant might add or delete one stroke, or change the direction of the stroke. For examples:孑孑,曾曾,黃黃,幵开

(2) Component :The component of variant may be added or deleted, or replaced by other one. For examples:匧篋,覓覕,珊姍,鍊煉,胃謂

(3) Position:The variant and glyph of missing character have the same components, but the arrangement is difference. Example:鵝鵞,蠏蟹,群羣

(4) Structure:The structure between variant and glyph is difference. Example:冊策,响響,祇只

The specification of MCDL is defined formally as follows:

```
<? xml version="1.0" ?>
<!ELEMENT    MCDL (Character+)>
<!ELEMENT    Character (Apply-glyph,Variants*) >
<!ELEMENT    Apply-glyph (Glyph-expression,
                    Code, Font, Difference? ) >
<!ELEMENT    Glyph-expression (#PCDATA) >
<!ELEMENT    Code (#PCDATA) >
<!ELEMENT    Font (#PCDATA) >
<!ELEMENT    Variants (Glyph-expression,
                    Difference? ) >
<!ELEMENT    Difference EMPTY>
<!ATTLIST    Code charset CDATA
                    #REQUIRED >
<!ATTLIST    Font typeface CDATA
                    #REQUIRED>
<!ATTLIST    Difference type (stroke | component|
                position | structure) "structure" >
```

The framework for interchanging missing character is shown in figure 3. There are some documents contain missing characters in the document set. Before exchanging documents, it is necessary to identify the missing characters in the document, then generates the MCDL description for missing characters by using ontology of Chinese etymology, finally, transmit documents and MCDL description to receivers.

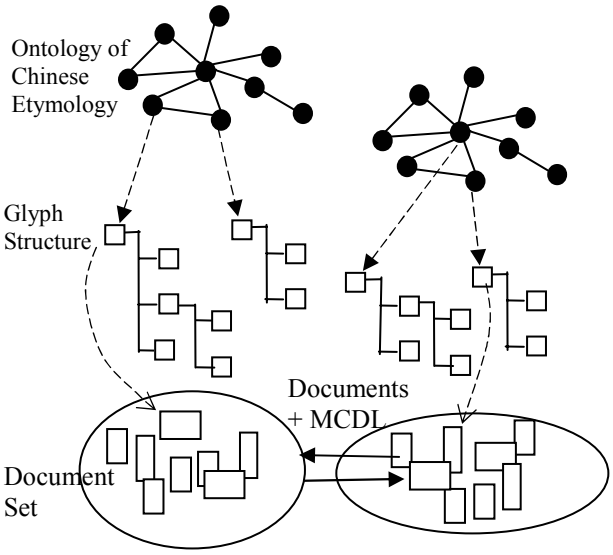

Figure 3. The framework for exchanging missing characters

4. The Design and Implementation

In this section, we briefly describe the six modules of interchange system for missing character that are currently implemented:

(1) Glyph structure database: this database is developed by Sinica[6], providing the glyph expression of 54000 glyphs of Hanyu Da Ziden to assist users in representing missing characters.

(2) Font database: if users need to add character in private use area, this database has kaishu font for 53693 glyphs.

(3) The ontology of Chinese etymology: this database contains about 5000 glyphs classified into 1100 categories, each contains variants for the same character. Each variant has modern, middle-ancient and ancient pronunciation. The middle-ancient pronunciation is based on Guanun and Giun. This ontology also describe the relation among variants to assist user choose proper variants to represent missing characters.

(4) Parser: It parses the text to identify the missing characters and generates the description using MCDL for each missing character. In addition, the parser processes the text received from unpacker to select properly glyph based on the MCDL and context environment of receiver.

(5) Packer: The packer will pack the original text file, the description and font for each missing character into a package prepared to interchange with other computers.

(6) unPacker:the unpacker unpacks the package received , then passes them to parser.

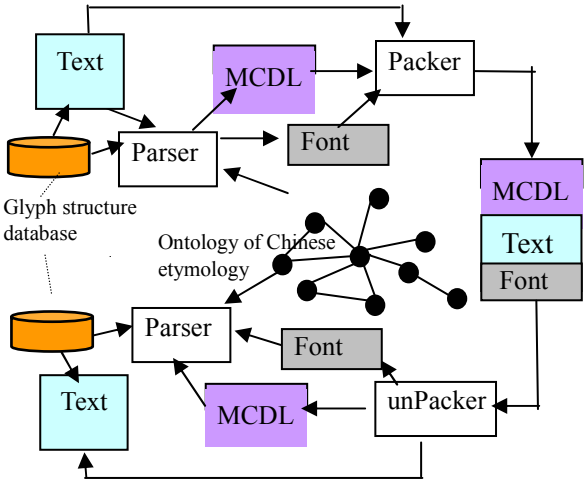

Figure 4. The design of interchange system

Based on our interchange framework, all documents have MCDL description for them. A metadata is also a document, so it has a MCDL description to describe the missing character in metadata.

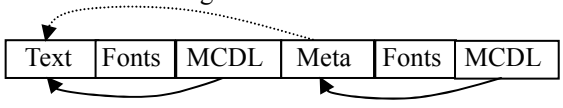

Figure 5. The combination of MCDL and metadata

To illustrate how the interchange system works, we take Chang A Han sutra in Taisho Tripitaka as an example. To digitalize Buddhist sutra, there are many characters can not find in any character set of encoding systems. Figure 6 shows the part of Chang A Han sutra contains three missing characters 鐵 (FA44,U+E004),驎(FA45,U+E005) and 積 (FA46, U+E006) in private use area.

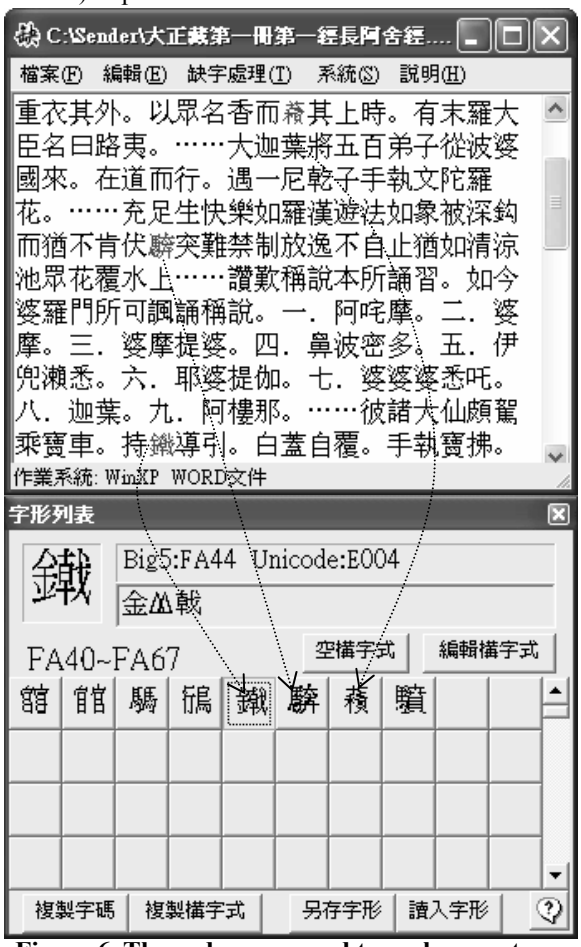

Figure 6. The sudra prepared to exchange at the sender.

The interchange system will automatically generate MCDL description for each missing character. 驎 has a variant 驥(as shown in figure 7), glyph expressions and the difference between them will be generated.

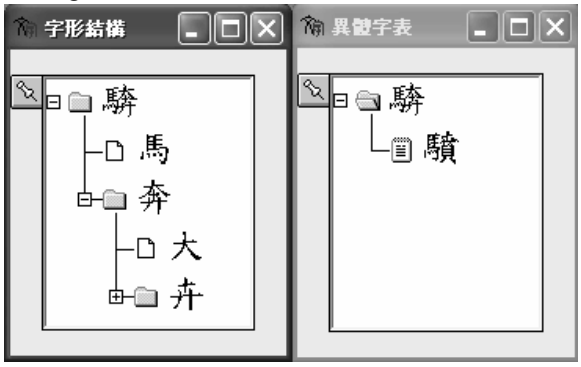

Figure 7. Glyph expression and variant for 驎

The complete description generated by interchange system is shown as follow:

```
<?xml version="1.0" encoding="big5" ?>
<mcdl:MCDL
        xmlns:mcdl="http://www.sinica.edu.tw/~cdp/
        missingchar-syntax/1.0">
 <mcdl:Character >
  <mcdl:Apply-glyph>
    <mcdl:Glyph-expression >
    馬厶奔</mcdl:glyph-expression>
    <mcdl:Code
    charset="big5">FA45</mcdl:code>
    <mcdl:Font   typeface="標楷體">FA45.gly
    </mcdl:Font>
  </mcdl:Apply-glyph>
  <mcdl:Variants>
   <mcdl:Glyph-expression >
   馬厶賁</mcdl:Glyph-expression>
   <mcdl:difference   type="component" />
  </mcdl:Variants>
 </mcdl:Character>
 <mcdl:Character >
  <mcdl:Apply-glyph>
   <mcdl:Glyph-expression >
   金厶戟</mcdl:Glyph-expression>
   <mcdl:Code charset="big5">FA44</mcdl:Code>
  <mcdl:Font   typeface="標楷體">FA44.gly
  </mcdl:Font>
  </mcdl:Apply-glyph>
 </mcdl:Character>
 <mcdl:Character >
  <mcdl:Apply-glyph>
   <mcdl:Glyph-expression>
   廾厽積</mcdl:Glyph-expression>
   <mcdl:Code charset="big5">FA46</mcdl:code>
   <mcdl:Font   typeface="標楷體">FA46.gly
   </mcdl:Font>
  </mcdl:Apply-glyph>
 </mcdl:Character>
</mcdl:MCDL>
```

At the receiver, 積 is already created in private use area with a different character code FA40(as shown in fig.8). The code FA44 is already assigned to different glyph. 驈 is not in private use area, but its variant is. If we transmit the Chang A Han sutra from sender to receiver, 鐵 will be replaced by 涳, 驈 replaced by 嚨, and 積 is shown as a blank. If we want to search 鐵,驈 or 積,this sutra will be not retrieved, causes the low recall. On the contrary, if we submit a query contains 涳 or 嚨, information retrieval system will deliver Chang A Han sutra to users, thus, causes the low precision.

By using the MCDL and interchange mechanism proposed in the paper, these problems all be solved. When the receiver gets the Chang A Han sutra, it immediately analyses MCDL to get the information about each missing character, and then compare the glyph expression with self-created characters in private use area of receiver. The receiver discovers that 積 has already existed and occupied the coding place at FA44. In addition, it has a variant 驥 to replace 驈. Because 鐵 does not exist in private use area of receiver, interchange system automatically adds 鐵 into private use area and allocate a unused character code. Figure 9 shows the result of private use area generated by interchange system. Because all missing characters have been properly processed by using MCDL, Chang A Han sutra can correctly be displayed and retrieved by information retrieval system as shown in figure 10 . Because the receiver has more information about missing character than before, it can properly process Chang A Han sutra.

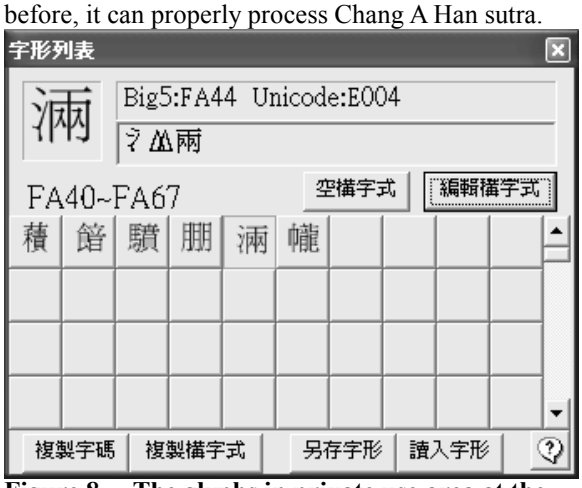

Figure 8. The glyphs in private use area at the receiver before interchange.

Figure 9. The glyphs in private use area at the receiver after interchange.

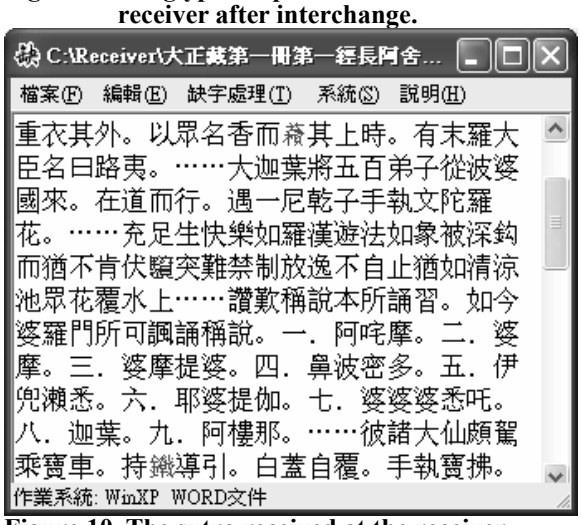

Figure 10. The sutra received at the receiver

5. Conclusions

In this paper, we develop a description language and framework to solve the interchange problem of missing characters. This approach has several advantages with respect to the methods accepted by previous studies:

(1) The methods of representing missing characters adopted by other applications can still be used without modification.
(2) MCDL can be easily integrated with RDFS, DAML+OIL and OWL which are languages for building ontology and semantic web.
(3) The character code of each missing character can be freely given by different users, because the identification of missing character depends on glyph expression not character code.
(4) The variants of missing character can be correctly exchanged and retrieved.
(5) Allowing users to add glyphs in private use area, it is very important because self-created glyphs are adopted by most people when they need the characters not in the character set.

However, the approach proposed in this paper does not consider the variants which can be replaced each other only in specific meaning. For example, both 雕 and 鵰 represent the concept of fearful birds, but, the concept of sculpture can only be represented by 雕. Therefore, it is necessary to further study the relationship among variants and describe them in the future.

References

[1]F. Coulmas, *Writing Systems:an Introduction to Their Linguistic Analysis*,2003.

[2]G. Tserdanelis and W.Y. Wong,*Language Files*,9th edition, 2004.

[3]C. Wittern and U. App."IRIZ Kanji Base : A New Strategy for Dealing with Missing Chinese Characters", *Electronic Buddhist Text Initiative,* April, 1996.

[4]C.Wittern,"A Proposal for TEI Character Encoding", *Workshop of Intelligent Encoding and Application for Han Characters,* March, 2003.

[5]C.C. Hsieh, "The Glyph and Encoding of Han Characters", *International Conference of Kanji code and database*, Tokyo, October, 1996.

[6]C.C Hsieh,"The Missing Character Problem of Electronic Ancient Books", *The 1st Chinese Etymology Conference,* August, 1996.

[7]T. Furuya, M. Maedera, H. Nomura, S.Tanimoto, and T.Yatagai, "How 90,000 Mjkyo Fonts are Working at Present by the Extension of UTF16", *Conference of Wen-Zi problems in Electronic Ancient Books*, June, 1999.

[8]J.S.Yen,"The Dynamic Generation of Han-Zi", *Workshop of Intelligent Encoding and Application for Han Characters,* March, 2003.

[9]D.Brickley and R.V.Guha, "Resource Description Framework Schema Specification". *W3C Proposed Recommendation*, March,2000.

[10]T. Berners-Lee, J.Hendler,and O.Lassila, "The Semantic Web. *Scientific American,* 2001.

[11]J.M.Chiou,"A Study on Missing Characters of Chinese Computer", Master Thesis, Hsuan Chuang University, 2001.

[12]K.J. Chen, "Computational Approaches in Topological and Geometrical Descriptions for Chinese Characters", *Computer Processing of Chinese & Oriental Languages, Vol. 2, No. 4,* October, 1986.

[13]Hanyu Da Zidian. 1st ed. Chendu: Sichuan Cishu Pulblishing, 1986

[14]Unicode version 4.0. http://www.unicode.org

[15]GB 18030: Xinxi Jishu-Xinxi Jiaohuan Yong Hanzi Bianna Zufuji-Jibenji De Kuochong. (Information Technology-Chinese Ideograms Coded Character Set for Information Interchange-Extension for the Basic Set). Beijing: Guojiao Zhilian Jishu Jianduju,2000.

[16]CNS11643:ZhongWen BiaoZhun JiaHuanMa (Chinese Standard Interchange Code), Taipei, 1992.

[17]CCCII:Zhongwen Zixun Jiaohuanma(Chinese Character Code for Information Interchange), Taipei:Xingzhenyuan Wenhua Jianshe Xiaozu (Executive Yuan Committee for Cultural Construction), Revised edition,1985.

[18]The Survey of Chinese Information Applications in Government, http://www.cns11643.gov.tw, 2003.

[19]X.G. Qiu, WenZiXue GaiYao (The Introduction to Chinese Etymology), 1995.