# Resolving the Unencoded Character Problem for Chinese Digital Libraries


## ABSTRACT

Constructing a Chinese digital library, especially for a historical article archiving, is often bothered by the small character sets supported by the current computer systems. This paper is aimed at resolving the unencoded character problem with a practical and composite approach for Chinese digital libraries. The proposed approach consists of the glyph expression model, the glyph structure database, and supporting tools. With this approach, the following problems can be resolved. First, the extensibility of Chinese characters can be preserved. Second, it would be as easy to generate, input, display, and search unencoded characters as existing ones. Third, it is compatible with existing coding schemes that most computers use.

This approach has been utilized by organizations and projects in various application domains including archeology, linguistics, ancient texts, calligraphy and paintings, and stone and bronze rubbings. For example, in Academia Sinica, a very large full-text database of ancient texts called Scripta Sinica has been created using this approach. The Union Catalog of National Digital Archives Project (NDAP) dealt with the unencoded characters encountered when merging the metadata of 12 different thematic domains from various organizations. Also, in Bronze Inscriptions Research Team (BIRT) of Academia Sinica, 3,459 Bronze Inscriptions were added, which is very helpful to the education and research in historic linguistics.



## Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.3.5 [**Information Storage and Retrieval**]: Online Information Services.

## General Terms

Languages

## Keywords

Character Encoding, Unencoded Chinese Characters, Digital Library, Glyph Expression


## 1. INTRODUCTION

Missing or unencoded characters for a language present practical problems in digital library applications including character encoding, text input, font generation and display, document distribution and searching, and so on. The problems are particularly serious for large character sets, such as the Chinese character[1] set, and other Han-related character sets. Constructing a Chinese digital library, especially for a historical article archiving, is often bothered by the small character sets supported by the current computer systems. This paper is aimed at resolving the unencoded character problem with a practical and composite approach for Chinese digital libraries.

In Chinese digital libraries, there's a fundamental problem in representing unencoded Chinese characters. The number of Chinese characters has been growing over time. This is due to the extensibility of Chinese characters which comes from the componential nature of characters. Moreover, the set of Chinese characters is not a closed finite set, and there's a lack of character standard.

The number of Chinese characters far exceeds the available coding space in computers. For example, the Simplified Chinese encoded in GB-2312 code standard contains 6,763 characters, while the Traditional Chinese in Big5 code standard contains 13,053 characters. Even though the Unicode Standard [15] tries to solve the problem by expanding the coding space to a much larger space of 70,207 characters in Unicode 3.1, the fundamental problem still exists: there're always some Chinese characters unencoded. According to a statistics in a full-text database for Chinese ancient documents called Scripta Sinica [14], there are 22,780 distinct characters in a total of 200 millions characters of ancient texts such as Ershiwu shi[2], Shisan jing[3], … etc. Among these 22,780 characters, 9,727 (42.7%) cannot be encoded using Big5 code. Even in modern usage of Chinese characters, the problem still exists. For example, some special characters in person names and location names cannot be correctly displayed and input. Due to naming convention or

[1] In this paper, we use the term "Chinese characters" to represent the CJKV ideographs or Han-related characters (漢字) which are commonly referred to as Hanzi in Chinese, Kanji in Japanese, or Hanja in Korean.

[2] Ershiwu shi 二十五史, The Twenty-five Dynastic Histories.

[3] Shisan jing 十三經, The Thirteen Confucian Classics.

personal preferences of characters, people can always coin new characters according to their needs. In fact, there are more than 18,000 unencoded Chinese characters in the person names of Household Registration in Taiwan. More than 10 new characters are being added per month.

Research of Chinese digital libraries started in the mid-1990s. In these digital libraries, various full-text databases of ancient and modern documents are being built. For example, the CDLP (China Digital Library Project) [8, 19] in China began in year 2000 with a 5-year timeframe building contents and infrastructures based on global standards. NDAP (National Digital Archives Project) [12] in Taiwan aims to promote and coordinate content digitization and preservation at leading museums, archives, universities, research institutes, and other content holders. During these efforts, four aspects of impact on these digital libraries were observed, including character encoding, text input, font generation and display, and document distribution and searching. First, in small-scale solutions, a "user-defined font area" in the coding space has been commonly adopted for users to specify a new code and the fonts for unencoded characters. However, the limited user space, the tedious font creation process, and the problem of code sequence ambiguity could make the digital library contents unable to be shared. Second, it would be very difficult to memorize and input the code sequences. Third, fonts have to be created by users and displayed when a user area code is detected. Finally, searching for such characters are difficult if not impossible. Therefore, a total solution for the unencoded Chinese character problem is required.

To resolve the unencoded character problem, there are three requirements. First, the extensibility of Chinese characters has to be preserved. Second, it should be as easy to generate, input, display, and search unencoded characters as existing ones. Third, it must be compatible with existing coding schemes that most computers use.

In this paper, a composite approach to the representation of Chinese characters is proposed. The proposed approach consists of the glyph expression model, the glyph structure database, and supporting tools. The glyph structure database contains 59,220 characters of Standard Scripts with 12,681 variants. Other characters include 177 Oracle Bone Inscriptions (甲骨文) with 25 variants, 3,459 Bronze Inscriptions (金文) with 1,055 variants, 11,100 Small Seal Inscriptions (小篆) with 1,081 variants, and 372 Chu Bamboo Slips Characters (楚系文字) with 73 variants. More characters are continuously added. The goal of this approach is to propose a complete solution that can fully express Chinese characters including existing ones (ancient and modern) and those to be coined. All these characters can be input, displayed, distributed, processed, and even searched in exactly the same way as ordinary characters without introducing too much inconvenience to users.

This approach has been utilized in various applications. First, in Academia Sinica, a full-text database of ancient texts called Scripta Sinica [14] were created which contains a very large collection of historic documents such as Ershiwu shi, Shisan jing, … etc. Second, the Union Catalog of NDAP tried to deal with the problem encountered when merging the metadata of 12 different thematic domains from various organizations. Third, in Bronze Inscriptions Research Team (BIRT) of Academia Sinica, 3,459 Bronze Inscriptions were added, which is very helpful to the education and research in historic linguistics.

# 2. BACKGROUND AND RELATED WORK

## 2.1 Chinese Characters and Glyphs

The Chinese writing system was originally pictographic, that is, words were represented by pictures that closely resembled the meaning of the word. The picture for "sun," for example, looked much like the sun. This pictographic writing eventually developed into the more complex ideographic writing that we are more familiar with. Chinese writing is one of the only contemporary writing systems that still prominently bear traces of its pictographic origins.

Around 1500BC, Chinese scripts started to spear in inscriptions. The first national standardization of Chinese scripts was done in about 200BC, in the realm of the First Qin Emperor (c. 259-210BC). Around 121AD, in the Eastern Han Dynasty, Shuowenjiezi[4] tried to analyze and classify all of the 10,516 characters at that time. In the 12th century, Guangyun[5] collected 25,126 characters. In 1716AD, Kangxi Dictionary[6] contains 47,043 characters.

Chinese characters are differentiated only by the meanings they carried. A character may have many glyphs or shapes. Glyphs are differentiated by their structures or skeletons. For example, the characters 'sun' and 'moon' have different meanings. The glyph of 'moon' in Standard Scripts looks quite different from the same character in Bone Oracle Inscriptions. During the history of Chinese characters, many different forms have been derived, for example, Oracle Bone Inscriptions, Bronze Inscriptions, Seal Inscriptions, Official Script, and Standard Script. The evolution of characters can help us trace the origin.

Glyphs do not deal with the actual outlook of the writing, but fonts do. A font is a collection of general design rules for a specific style of typefaces. For example, it's possible to specify the size, the broadness of strokes, the ratio of broadness of vertical and horizontal strokes, etc. For the same glyph, different styles of writing and calligraphic variants such as 宋 (Song), 隸 (Clerical Scripts), 行 (Running Scripts), 草 (Grass Scripts), etc. can also exist.

Chinese characters are extensible or productive in the sense that new characters can be generated out of old ones. It's fundamental since a finite set of basic ideographic components and composition operators can be used to express an unlimited set of unique Chinese characters. There are three types of elements in Chinese characters: strokes, components, radicals,.

---

[4] Shuowenjiezi 說文解字, "Elucidations of the Signs and Explications of the Graphs," compiled by Xu Shen (許慎), 100AD.

[5] Guangyun 廣韻, an ancient Chinese rhyme dictionary.

[6] Kangxi Dictionary 康熙字典, a dictionary compiled by Zhang Yushu (張玉書) et. al., 1716AD.

Strokes are the atomic element constituting the shape in Chinese characters. It is the basic element in writing. From a serial group of strokes, components can be formed. There are two types of components: basic and combined. Basic components, also known as radicals, are the minimum fundamental components that will not be further decomposed. Combined components are intermediate elements consisting of several radicals. In order to uniquely represent a glyph, it's natural to decompose a glyph in terms of its radicals.

In 1972, Ni [13] analyzed 16 different compositional operators of Chinese glyphs and found that only three of them, namely horizontal composition, vertical composition and contain composition, are frequently used and can be assembled as an effective system for representing the structure of glyphs. After analyzing the 9,129 glyphs in [7], 496 radicals were obtained. Named as the NCTU Radical Set (a.k.a Chiao-Tung Root System), it was the earliest study of glyph structure in Taiwan.

This radical set is unique in that it is obtained by optimizing the total number of the radicals and the averaged number of radicals per glyph. In general, the less the number of radicals, the longer the decomposition of glyph. The weighted average number of radicals per glyph is only 1.9. After checking against the 49,905 characters of a Chinese lexicon[7], 48,713 characters can be expressed by this set. The remaining 1,192 characters are some ancient scripts. More details of the statistics and optimization of the NCTU Radical Set are illustrated in the Appendix.

## 2.2 Chinese Character Encoding Schemes

In order to represent Chinese characters in computers, many different encoding schemes have been proposed to distinguish among characters. For example, unique 2-byte codes such as GB-2312, Big5, …, etc. were assigned to each character. However, the encoding standard for Chinese characters is still in lack. In China, the Simplified Chinese encoded in GB-2312 code standard contains 6,763 characters, while GBK code contains 20,902 characters. In Taiwan and Hong Kong, the Traditional Chinese in Big5 code standard contains 13,053 characters, while the CNS 11643 code standard contains 48,027 characters. The Unicode Standard 3.1 [15] tried to unify all CJKV ideographs with a set of 70,207 characters.

Since there are more than 65,536 unique Chinese characters, only the frequently used characters can be represented if each character is represented by one 2-byte code. In existing encoding standards such as GB-2312 and Big5, no extra information about characters is encoded except the code sequence of glyphs, which has no particular meaning besides indexing. In the character-glyph model [15], the underlying model in Unicode, abstract characters with semantic meanings can be differentiated from concrete glyphs, the visual shapes of characters. Only characters are encoded, not glyphs [3]. Characters from existing CJKV character sets must be merged, not separately encoded. Therefore, the largest subset of the Unicode Standard, the Unihan database [6] was introduced which was the repository for the Unicode Consortium's knowledge regarding CJKV ideographs.

Different from the Unicode model, we adopt an analytic representation of characters, by breaking them down into pieces or components. This is similar to an ideographic composition scheme called Ideographic Description Sequence (IDS) [9, 10] later adopted by Unicode with different set of operators and components. Recently, a character description language (CDL) [1] was also proposed to describe the characteristics of characters, which is more versatile than IDS.

## 2.3 International Efforts in Chinese Digital Libraries

There are many international efforts in dealing with the problems encountered when building Chinese digital libraries. Wittern [17, 18] introduced the problems encountered when dealing with digitization of Chinese Buddhist texts. CHANT project [4] in Hong Kong aims to build an electronic database of all 1500BC to 600AD Chinese ancient texts. Richard Cook [2] discussed the character set issues relating to the computerization of an ancient Chinese lexicon Shuowenjiezi.

In China, the CDLP (China Digital Library Project) [8, 19] was a 5-year project starting from year 2000 building contents and infrastructures based on international standards. In Taiwan, NDAP (National Digital Archives Project) [12] starting from 2002 aims to digitize all kinds of databases across various organizations.

# 3. GLYPH EXPRESSION AND GLYPH STRUCTURE DATABASE

In the proposed approach, there are three components, the glyph expression, the glyph structure database, and supporting tools.

## 3.1 Representation of Chinese Glyphs

Our glyph model is shown in Table 1. In this model, there are two parts: the representation of glyphs and the glyph structure expression. A formal representation in Backus Naur Form is shown in Table 1.

**Table 1. A formal representation of Chinese glyphs in Backus Naur Form.**

| | |
|---|---|
| <character set> ::= | <glyph> \| <symbols> |
| <symbols> ::= | punctuation symbols, pronunciation symbols, and others |
| <glyph> ::= | <radical> \| <component> \| <glyph><operator><glyph> |
| <component> ::= | <radical> \| <component><operator><component> |
| <operator> ::= | - (horizontal) \| ^ (vertical) \| @ (containing) |
| <radical> ::= | 492 radicals |

Mathematically speaking, the representation in Table 1 is productive in that the possible produced outcomes, such as character, glyph and component are usually far more than we

[7] Zhongwen da cidian 中文大辭典, "Great Dictionary of the Chinese Language," edited by Zhang Qiyun, 40 vols., 1962-1968. Revised edition, 1973, 10 vols.

may accept. Whether the outcome is a legal item or not is up to our choice. Thus, this representation has the property of extensibility that just fit the requirement.

In this productive system, the structure of a glyph is expressed as an expression of radicals. Radical is a basic component which will not be decomposed further. An example of glyph structure decomposition is shown in Figure 1. The term component is usually referred to the intermediate elements between a glyph and its radicals. In Figure 1, 藏 and 戕 are glyphs, 戈 is a component, 艹, 爿, and 臣 are radicals, where 臣 is also a glyph.

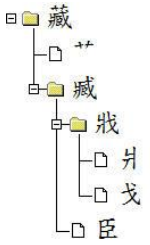


**Figure 1. An illustration showing the glyph decomposition of a character "藏".**

Example sets of radicals and components are shown in Figure 2 and Figure 3, respectively.

**Figure 2. An example set of radicals.**

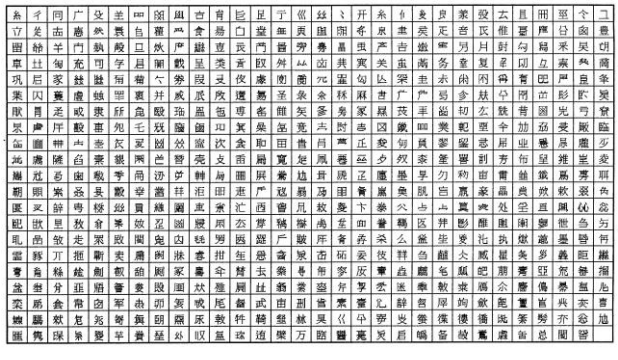

**Figure 3. An example set of components.**

## 3.2 Glyph Structure Expression

*Glyph structure expression* (or *glyph expression* for short) can be used to represent any existing character and those to be created in a user friendly way, which helps people memorize the code sequence very easily. Existing characters including ancient and modern characters can be looked up using the glyph expression. Ancient scripts such as Oracle Bone Inscriptions, Bronze Inscriptions, Seal Inscriptions and their evolutionary relationships can also be represented.

For any new unencoded character to be coined, they can also be created and represented in the same way as existing characters. All one need to know is the shape and components of a glyph.

Input methods like Changjie (倉頡) also uses the components to describe the character. In designing an input method, there is a tradeoff between the length of the code sequence representing a character and the number of conflicting characters with exactly the same code sequence. For users, it's important that a code sequence is efficient to type and easy to remember. The number of keystrokes (and thus the length of code sequence) should be minimized. For coding efficiency, the longer the code sequence, the more characters that can be represented, and also the less conflicting characters.

Different from an input method like Changjie, we are not concerned about the number of keystrokes. Instead, the extensibility of characters and the user convenience for looking up characters and creating new ones are the primary consideration. Therefore, glyph expression can represent the most characters with a very small percentage of conflicting representations.

Let us explain how the structure of a glyph is expressed by examples. In Figure 1, the glyph '藏' can be decomposed as:

藏 = 艹 ^ 臧
臧 = 戕 @ 臣
戕 = 爿 – 戈 (1)

In Equation 1, these formulae are generally referred to as *glyph expressions*, in which "^", "-", and "@" represent the operators of vertical, horizontal, and contain compositions, respectively. An expression consisting of only radicals can be obtained by successively decomposing components as follows:

藏 = 艹 ^ 臧
= 艹 ^ (戕 @ 臣)
= 艹 ^ ((爿 – 戈) @ 臣) (2)

In Equation 2, the glyph "藏" is composed of 4 radicals. Note that these 4 radicals are all included in Big5 code. Such glyph expression consisting solely of radicals is called a *radical expression*. Similarly, a glyph expression by components such as Equation 1 may be called a *component expression*. When all the operators are eliminated, equation 2 becomes:

藏 = 艹 爿 戈 臣 (3)

Equation 3 is called the *radical sequence* of the glyph 藏 showing the compositional order of radicals. Following the same naming thought, the formulae in Equation 1 after eliminating operators may be called *component sequences* which show the compositional order of components.

In general, any representation called "expression" refers to a complete glyph structure information, while "sequence" does not. Each glyph has a unique glyph expression and hence this expression can be served as the identifier of that glyph.

Although “sequence” does not have complete structure information of glyph, it still has very high discriminating abilities among glyphs. For instance, in the 9,129 glyphs of [7], there are only 8 pairs of glyphs with exactly the same component sequence, such as （唄, 員）. All others can be uniquely identified by their component sequences.

According to our analysis, very few characters have exactly the same glyph expressions, for example, (稽, 尤旨 ). There are also some characters with the same sequence, but different expression, for example, (峰, 峯). This shows the uniqueness of glyph expression and also the very low percentage of conflicting characters. Since the representation is natural and intuitive in terms of character writing and composition, people can easily learn to use the glyph expression to represent characters in a short period of time. It’s more convenient than user-defined font area since less laborious work needed for users.

## 3.3 Chinese Glyph Structure Database

Based on the idea of glyph expression model, it is easy to implement the glyph structure database (or glyph database for short), which contains the knowledge representation of glyph structure information for Chinese characters. For an unencoded character, we can look it up in the glyph database by its glyph expression or from the related characters by component lookup in the user interface of the glyph database as shown in Figure 4. We can easily input an unencoded character in a document. It can also help create characters including ancient ones. With the organization of glyph knowledge, the glyph database is more powerful than a font table.

This database has been continuously maintained for many years. The current scale of the character set was enlarged from the set in [7] by including Simplified Chinese characters and the characters in Zhongwen da cidian and Hanyu da zidian[8]. As of the writing of this paper, the glyph structure database contains 59,220 characters of Standard Scripts with 12,681 variants. Other characters include 177 Oracle Bone Inscriptions with 25 variants, 3,459 Bronze Inscriptions with 1,055 variants, 11,100 Small Seal Inscriptions with 1,081 variants, and 372 Chu Bamboo Slips Characters with 73 variants. These ancient characters are precious resources, and more characters are being added.

Containing the rich knowledge of character composition, character evolution, and relational mappings among different forms of characters, the glyph database also provides an integrated user interface for exploring and searching within the database. As shown in Figure 4, both ancient and modern characters can be looked up by components, radicals, or glyph expressions within the glyph database. Then, the components and their structures can be displayed. In addition, the variants, the evolution, and the relations among various forms of the characters can be displayed. The related source information such as Shuowenjiezi can also be looked up.

[8] Hanyu da zidian 漢語大字典, Great Dictionary of Chinese Characters, edited by Hanyu da zidian weiyuanhui. Wuhan: Hubei cishu chubanshe and Sichuan cishu chubanshe, 1986..

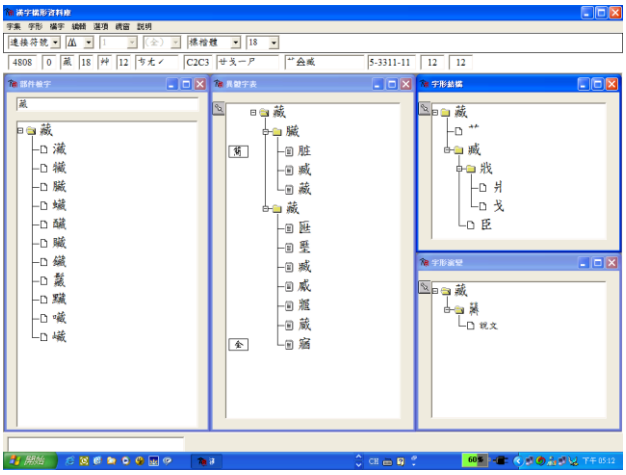

**Figure 4. An illustration showing the integrated user interface for the glyph database.**

### *3.3.1 Optimization to the Glyph Database - the Kernel Set*

There’s a tradeoff between the user-friendliness of glyph expressions and the total number of components in the glyph database. The total number of components in the glyph database has to be minimized for encoding space efficiency. On the other hand, to maintain an easy-to-use glyph expression, the average number of components per glyph has to be minimized. Therefore, there’s a need for an optimization to the glyph database. This can be done by limiting the operators in a component expression to be one. In this case the number of components per glyph is usually 2, and occasionally 3. After all one-operator expressions of a character set have been collected, a kernel set which has all necessary elements for one-operator expression for any one character/glyph in the original character set can be found. A pictorial presentation of the kernel set of our glyph database is shown in Figure 5. This kernel set includes 457 radicals (in which 269 are also characters), 629 components, and 1,044 kernel glyphs. There are 2,130 elements in total, where 1,313 are characters and the total frequency of usage is 62.60%.

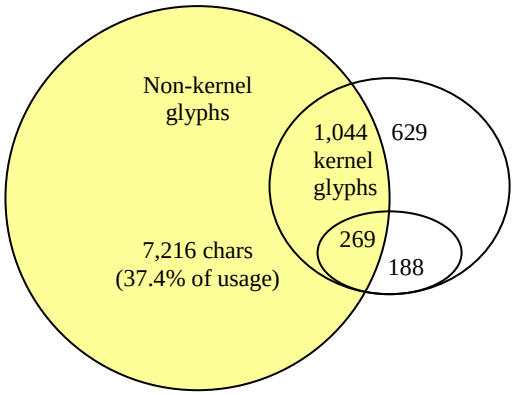


**Figure 5. An illustration showing the kernel set of glyph database.**

The kernel set in Figure 6 is not very efficient in terms of the total frequency of usage (only 62.60%). This can be optimized by including frequently used characters into the kernel set. A frequency distribution chart of our glyph data base is shown in Figure 6. In Figure 6, it is obvious that including the 1,071 characters of the most frequently used category into the kernel will raise the total frequency of usage of the kernel to 97.8%. The price paid is the increase of the numerical coding space to 3,201. It is not bad and can be afforded by any 2-byte coding scheme.

Following the same thought, there are 5 levels of possible optimization of the kernel set as listed in Table 2. In Table 2, the fourth choice is recommended which includes all glyphs of the most frequently used and the frequently used categories. It contains 4,946 elements and the averaged code length is only 1.007 times of a 2-byte code.

**Table 2. The weighted average code length at different levels of optimization in the character sets.**

| Character Set | Number of Codes | The Weighted Average Code Length | Entropy |
|---|---|---|---|
| 1. Radicals | 457 | 1.9+1=2.9 | 7.3038 |
| 2. Kernel Set | 2,130 | (37.4*3+62.6)%=1.748 | 8.6782 |
| 3. Kernel Set & Most Frequently Used Characters | 3,201 | (2.2*3+97.8)%=1.044 | |
| 4. Kernel Set & Frequently Used Characters | 4,946 | (0.368x*3+99.632)% =1.007 | |
| 5. The whole set | 9,346 | 1 | 9.1982 |

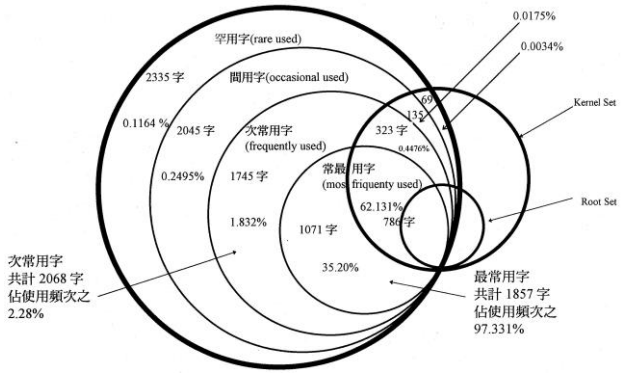


**Figure 6. An illustration showing the kernel set and the frequency distribution chart.**

### *3.3.2 A Descriptive Method of Coding*

To save the encoding space, a descriptive method of coding is adopted for characters not included in the kernel set. In ordinary numeric coding, each character is assigned a numeric code. This works for characters in the kernel set since it is a closed set with a relatively small (2,130) total number of elements. They can be easily coded into a 2-byte code space. For characters not included in the kernel set, descriptive coding can be used where glyph expressions are used as the code words for characters instead of ordinary numeric codes. Descriptive code word does not require ordinary numerical coding space. Besides, descriptive coding is a productive system which is capable of representing additional newly created characters/glyphs to the existing character set.

Descriptive coding is compatible with existing codes. When 629 components and 188 non-character radicals of the kernel we developed as shown in Figure 5 are included into an existing coding scheme such as GB, Big5, or Unicode, the descriptive capability can be provided. It can be further optimized by excluding some rarely used characters to save coding space, if necessary. If a kernel set for each country can be derived, then a CJK unified kernel can be formed, and hence a CJK unified descriptive code can be constructed.

## 4. Supporting Tools

In order to support the glyph expression and glyph database, tools for character encoding, text input & display, font generation, document searching are required.

A scenario of the lifecycle of processing documents with unencoded characters is illustrated in Figure 7.

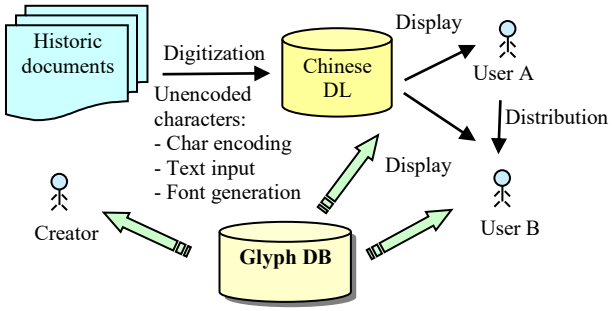


**Figure 7. An illustration showing a scenario of the lifecycle of documents containing unencoded Chinese characters.**

As shown in Figure 7, the creator of a Chinese digital library has to first digitize the historic documents or antiques such as bronze and stone rubbings. In the digitization process, several issues arise for unencoded characters, for example: character encoding, text input, and font generation, among others. Second, for a user to see the digitized objects, a display tool has to be provided for the unencoded characters to be correctly shown. Third, when User A wants to distribute or disseminate the documents in Chinese digital libraries to User B, the information about the unencoded characters must be carried so that the same characters can be correctly input, displayed, and further utilized. Finally, applications such as document searching can then be supported.

To support the processing of documents containing unencoded Chinese characters with the glyph database, the necessary modules are illustrated in Figure 8.

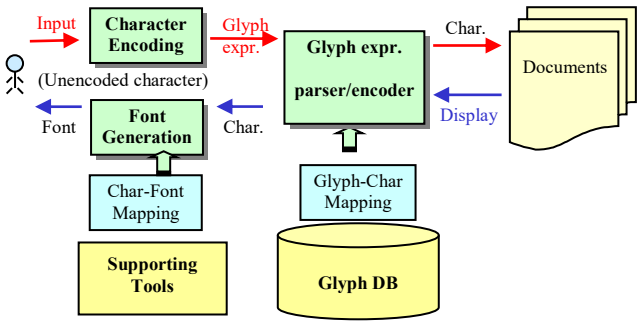


**Figure 8. An illustration showing the components in supporting tools for unencoded Chinese characters.**

## 4.1 Tools for Character Encoding and Text Input

As shown in Figure 8, there are two components when inputting texts: character encoding and glyph expression/markup tag processing. In order to represent unencoded characters in ordinary documents, a character encoding module is needed getting the glyph expression for unencoded characters. The glyph expression of an unencoded character can be obtained either by inputting glyph expressions directly or by component look up from the user interface of glyph database. Then, a glyph expression parser/encoder module is needed to convert glyph expressions into character representation in documents. Markup tags are needed to indicate the beginning and ending of the code sequence and glyph expression.

The user is not required to look for the missing glyph elsewhere in order to obtain an identifier or a code word for the unencoded character, a database of collecting various glyphs and variants is not required at the user's front end, and finally, the relational mapping between character and glyph can be fully expressed.

There are several advantages to using glyph structure in coding scheme. Firstly, the component set as well as the radical set of Chinese characters is a closed system. It will not expand indefinitely as the character set does. Although there are rare cases when they might need to be extended, they are far more manageable than that of character set.

The second advantage is that the glyph structure model is a productive system, as described earlier. This means it has the extensibility and flexibility of not changing the existing system over newly created characters. Thirdly, glyph structure is a kind of knowledge representation. The character code can be represented by glyph expression or component sequence. It not only facilitates character coding and human reading, but also embeds more knowledge about characters into the existing coding scheme for further applications in Chinese information processing.

Since the glyph expression is unique for most glyphs, it can be used as the identifiers for characters. To input a character, the glyph expression can be directly input and the corresponding character code will be generated by glyph expression parser/encoder module. Alternatively, users can input components or radicals from the user interface of glyph database where relevant glyph expressions of similar characters can be found. If there's any ambiguity, different alternatives can be shown for the user to choose. Since the uniqueness of the glyph expression as previously mentioned, only very few characters have exactly the same glyph expression.

## 4.2 Tools for Font Generation and Display

For a character to be correctly displayed, it has to be first converted from glyph expressions. Then, its associated font has to be generated and displayed. A resource locator indicating the location of the corresponding fonts will be encoded in texts.

For a document with glyph expressions, the glyph expression parser is needed to parse the expressions into corresponding characters and a font generation module is used to get the corresponding fonts for glyphs in glyph database. In fact, glyph database served as the index for glyph-character mapping and character-font mapping for supporting tools.

## 4.3 Tools for Document Distribution and Searching

Since the glyph expression is a unique representation of characters, it's natural to distribute the documents in glyph expressions instead of the converted form of glyph fonts or pictures. No user-defined fonts are needed. Also, the embedded information about the characters includes the locators of the fonts in glyph database. Information exchange can be done in a more natural and smooth way in which unencoded character problem can be effectively resolved.

As long as the glyph expression is used as the identifier of characters, the unencoded characters in glyph expressions can be searched in the same way as normal characters. Searching of a character can be done by glyph expression, components, or radicals. It works in the same way as character input does.

# 5. APPLICATIONS IN DIGITAL LIBRARIES AND DISCUSSION

Currently, the glyph expression and glyph database have been utilized by organizations and projects in various areas including archeology, linguistics, ancient texts, calligraphy and paintings, and stone and bronze rubbings (金石拓片), to name a few.

For example, in Academia Sinica, a full-text database for Chinese ancient documents called Scripta Sinica [14] contains more than 200 million characters of ancient texts such as Ershiwu shi, Shisan jing, … etc. More than 9,727 characters are unencoded if the documents were to be represented in Big5 code. Only 427 of them cannot be directly represented by our approach. That is, 9,500 (more than 95.6%) of them can be directly resolved without adding new radicals or components.

In NDAP [12], there are currently 9 public institutions participating in the project. These include nearly 100 different databases from libraries, museums, universities, and research institutes in 12 thematic groups such as geology, zoology, botany, anthropology, archives, calligraphy & painting, and rare books. In order to integrate the metadata and the mappings among various databases from such diverse domains, an OAI-based [11] Union Catalog was developed [16]. Currently, 140,000 pieces from the Institute of History and Philology of

Academia Sinica (IHP)[9] and 10,000 pieces from Han Wooden Slips (漢代簡牘資料庫) are being integrated. The unencoded character problem was among the problems encountered. Glyph database was also utilized to resolve the problem.

In Bronze Inscriptions Research Team (BIRT, 金文工作室) under IHP, a Lexicon of Pre-Qin Bronze Inscriptions of Bamboo Scripts (先秦金文簡牘詞彙資料庫) has been constructed where 3,435 characters are unencoded using Big5 code. Among them, 1,462 characters (42.6%) can be directly represented using our approach. The remaining 1,973 unencoded characters together with 2,253 more characters collected from other references on Bronze Inscriptions were added into the glyph database.

As of this writing, the Dec. 2004 version of the glyph database was announced and included in a CD-ROM. The database, documents, papers, and related information are also available for download in the following website: http://www.sinica.edu.tw/~cdp/.

### *Results*

As an example result, the original message of a short passage (in Bronze Inscriptions) scripted on a Bronze ritual vessel is shown in Figure 9 together with the message in Big5 code, and the message in glyph expression.

(a)


(b)

“白（伯）父乍（作）／寶鼎，其子子／孫孫永用。〔幷〕。”

(c)

“白（伯）章　𠂤父乍（作）／寶鼎，其子子／孫孫永用。〔幷〕。”

**Figure 9. An illustration showing the short passage scripted on a Bronze ritual vessel (伯章　𠂤父鼎): (a) the original image, (b) the message in Big5 code, and (c) the message in glyph expression.**

As shown in Figure 9, the unencoded character “章　𠂤“ can be correctly displayed in glyph expression. More examples are available upon request.

[9] IHP (the Institute of History and Philology of Academia Sinica), http://www.ihp.sinica.edu.tw/.

### *Discussion*

User-defined font area or private user area is only a small-scale short-term solution to the unencoded character problem. It’s neither scalable, nor portable.

On the other hand, our approach identifies each character with its glyph expression, which is extensible to large-scale full-text Chinese databases or digital library applications. Also, the font is stored as a separate resource on the Web, and only resource locators are required to be embedded in documents. The distributed approach helps resolve the portability issue.

The glyph expression model is extensible in that a finite set of components can be used to derive an open set of characters, including all possible forms from the past and those to be coined in the future.

The current glyph database is based on the extension to existing Big5 coding space. The integration with other coding schemes such as Unicode is in progress.

## 6. CONCLUSION

In this paper, we proposed a glyph expression model for unencoded Chinese characters in digital libraries. The glyph expression model and the glyph database were introduced to resolve the unencdoed character problem. A scenario of dealing with documents containing unenecoded character was illustrated as well as the supporting tools for applying the model in character encoding, font generation, text display and input, and document dissemination. The statistics on current applications in Academia Sinica, NDAP program, and research teams on various domains also showed its potential in more applications in Chinese digital libraries.

## 7. APPENDIX

Lin [7] conducted a research on finding “A Fundamental Chinese Character Set for Computer Use” aiming at all possible glyphs that computer might possibly confront, which is different from all previous works. A thorough survey of 11 character sets used by popular dictionaries and press media during that period had been done. This covers more than 30 previous statistical studies on Chinese characters from 1856 to 1971.

This research collected 8,532 characters with additional 597 variants. The total frequency count of usage is 2,022,604, and their distribution together with their statistics is listed as follows. After some corrections to this character set, 9,122 glyphs are obtained, which includes 8,529 characters and 593 variants. Consisting of 629 components and 457 radicals, this character set contains 9,751 glyph expressions, which forms the basis of this work.

Up to now, this survey is still the most comprehensive one and we use it as a foundation of our study. Table 3 shows the number of characters and their percentage of use in [7].

**Table 3. An illustration showing the number of characters and their percentage of use in [7].**

| Category | Number of Characters | Percentage of Use |
|---|---|---|
| Most Frequently Used | 1857 | 97.34% |
| Frequently Used | 2068 | 2.27% |
| Occasional Used | 2182 | 0.27% |
| Rarely Used | 2425 | 0.12% |
| Total | 8,532 | 100.00% |

The entropy of the system is 9.60. More details on the statistics and optimization of the NCTU Radical Set can be found in [5].

## 8. REFERENCES


[1] Bishop, T. and Cook, R.S. A Specification for CDL Character Description Language. *In Glyph and Typesetting Workshop*, Kyoto, Japan, 2003.

[2] Cook, R.S. The Extreme of Typographic Complexity: Character Set Issues Relating to Computerization of the Eastern Han Chinese Lexicon Shuowenjiezi. In *Proc. of the 18th International Unicode Conference (IUC-18)*, Apr. 2001.

[3] Cook, R.S. Typological Encoding of Chinese: Characters, Not Glyphs. In *Proc. of 19th International Unicode Conference (IUC-19)*, Sep. 2001.

[4] Ho, Che Wah. CHANT (CHinese ANcient Texts): a Comprehensive Database of All Ancient Chinese Texts up to 600 AD, *Journal of Digital Information, Volume 3 Issue 2*, Article No. 119, Aug. 2002.

[5] Hsieh, C.C (謝清俊). A Descriptive Method for Re-engineering Hanzi Information Interchange Codes. Japan, Oct. 1996. http://www.sinica.edu.tw/~cdp/paper/1996/19961005_1.htm.

[6] Jenkins, J.H. The Dao of Unihan. In *Proc. of the 26th International Unicode Conference (IUC-26)*, Sep. 2004.

[7] Lin, Su (林樹). Research on the Fundamental Chinese Character Set for Computer Use (中文電腦基本用字研究), Technical Report, Department of Computer and Control Engineering, NCTU, March 1972. (in Chinese)

[8] Liu, Wei. The Development of Digital Collections and Metadata Applications in Chinese Libraries. In *Proc. of International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society (DLKC 2004)*, Japan, Mar. 2004.

[9] Lu, Q (陸勤). The Ideographic Composition Scheme and Its Applications in Chinese Text Processing. In *Proc. of the 18th International Unicode Conference (IUC-18)*, Apr. 2001.

[10] Lu, Q., Chan, S., Li, Y., and Li, N. Decomposition for ISO/IEC 10646 Ideographic Characters. In *the 3rd Workshop on Asian Language Resources and International Standardization, COLING 2002*, Taipei, 2002.

[11] OAI (Open Archives Initiative), http://www.openarchives.org/.

[12] NDAP, National Digital Archives Program, Academia Sinica (http://www.ndap.org.tw/)

[13] Ni, Keng (倪耿). 中國文字之結構模式及其分析, Master's Thesis, Institute of Electronics, NCTU, 1972. (in Chinese)

[14] Scripta Sinica, Hanji dianzi wenxian 漢籍電子文獻, Academia Sinica, http://www.sinica.edu.tw/~tdbproj/handy1/.

[15] The Unicode Consortium. The Unicode Standard, Version 4.0.1, defined by: *The Unicode Standard, Version 4.0* (Reading, MA, Addison-Wesley, 2003. ISBN 0-321-18578-1), as amended by Unicode 4.0.1 (http://www.unicode.org/versions/Unicode4.0.1/)

[16] Union Catalog of NDAP, http://catalog.ndap.org.tw/.

[17] Wittern, Christian, Chinese Buddhist texts for the new Millennium — The Chinese Buddhist Electronic Text Association (CBETA) and its Digital Tripitaka, *Journal of Digital Information, Volume 3, Issue 2*, Article No. 123, Sep. 2002.

[18] Wittern, Christian and App, Urs, IRIZ Kanji Base: A New Strategy for Dealing with Missing Chinese Characters, *EBTI (The Electronic Buddhist Text Initiative)*, Taipei, April 1996.

[19] Yang, G. and Zhang, T. The Development of the China Digital Library. *Electronic Journal of Academic and Special Librarianship, Vol. 3, No. 3*, 2002.