

A Mechanism for Solving the Unencoded Chinese Character Problem on the Web

Te-Jun Lin¹, Jyun-Wei Huang², Christine Lin², Hung-Yi Li²,
Hsiang-An Wang¹, Chih-Yi Chiu¹

¹Institute of Information Science, Academia Sinica, Taiwan

²Dept. of Information Management, Yuan Ze University

¹{soar, sawang, cychiu}@iis.sinica.edu.tw,

²{s932658, s941714, s941712}@mail.yzu.edu.tw

Abstract. The unencoded Chinese character problem that occurs when digitizing historical Chinese documents makes digital archiving difficult. Expanding the character coding space, such as by using the Unicode Standard, does not solve the problem completely due to the extensibility of Chinese characters. In this paper, we propose a mechanism based on a Chinese glyph structure database, which contains glyph expressions that represent the composition of Chinese characters. Users can search for Chinese characters through our web interface and browse the search results. Each Chinese character can be embedded in a web document using a specific Java Script code. When the web document is opened, the Java Script code will load the image of the Chinese character in an appropriate font size for display. Even if the Chinese characters are not available in the database, their images can be generated through the dynamic character composition function. As the proposed mechanism is cross-platform, users can easily access unencoded Chinese characters without installing any additional font files in their personal computers. A demonstration system is available at <http://char.ndap.org.tw>.

Keywords: Chinese glyph structure database, digital archive, unencoded Chinese characters

1 Introduction

The unencoded Chinese character problem often causes difficulties when historical Chinese documents are digitized, as there are always some unencoded Chinese characters in them. Expanding the character coding space, such as by using the Unicode Standard, does not solve the problem completely due to the extensibility of Chinese characters. User can create new characters by reshaping or composing existing characters to meet their needs. For example, the Chinese character, "王" ("the king"), may have many different forms in Oracle Bone Inscriptions, "𠩺" (甲骨文), Bronze Inscriptions, "𠩺" (金文), and Seal Inscription "𠩺" (篆書), respectively. To correctly display and input these unencoded characters, conventional approaches require the installation of additional font files in personal computers. To address the

problem, Juang *et al.* [2][3] developed a Chinese glyph structure database to facilitate searching and generating unencoded characters.
















In this article, we propose a mechanism that extends the capability of Juang's database to solve the unencoded Chinese character problem encountered in Web applications. We provide a web-based interface to search for Chinese characters, each of which can be embedded in a web document with a specific Java Script code. When the web document is opened, the embedded code will load the image of the character in an appropriate font size for display. If some characters do not exist in the database, their images can be generated dynamically through our character composition function. As the proposed mechanism is cross-platform, users can easily access and browse unencoded Chinese characters in web documents without installing any additional font files in their personal computers.

2 Methods and Techniques

Since the proposed mechanism is based on Juang's Chinese glyph structure database, we begin with an overview of the database, which currently contains 115,197 Chinese characters. A glyph expression is used to represent a Chinese character code. The database defines three categories of glyph expressions that cover thirteen "glyph operators," as shown in Table 1. The Chinese glyph structure database can be installed in personal computers and integrated with Microsoft Word.

The proposed mechanism is illustrated in Fig. 1. Below, we describe the mechanism's three major components, namely the Unencoded Character Retrieval System, the Chinese Glyph Expression API, and the Dynamic Character Composition Function.

Table 1. Glyph Expressions

Category	Operator	Explanation
Connection		The components are connected from left to right.
		The components are connected from top to bottom.
		The components are connected from outside to inside.
Component Sequence		Input connecting components in sequence. Add the start operator () first and the end operator () last.
		
Convenient Operator		The two components are the same and connected vertically.
		The three components are the same and connected vertically.
		The two components are the same and connected horizontally.
		The three components are the same and connected horizontally.
		The three components are the same and arranged as a triangle.
		The four components are the same and connected horizontally.
		The four components are the same and connected vertically.
		The four components are the same and arranged as a tetragon.

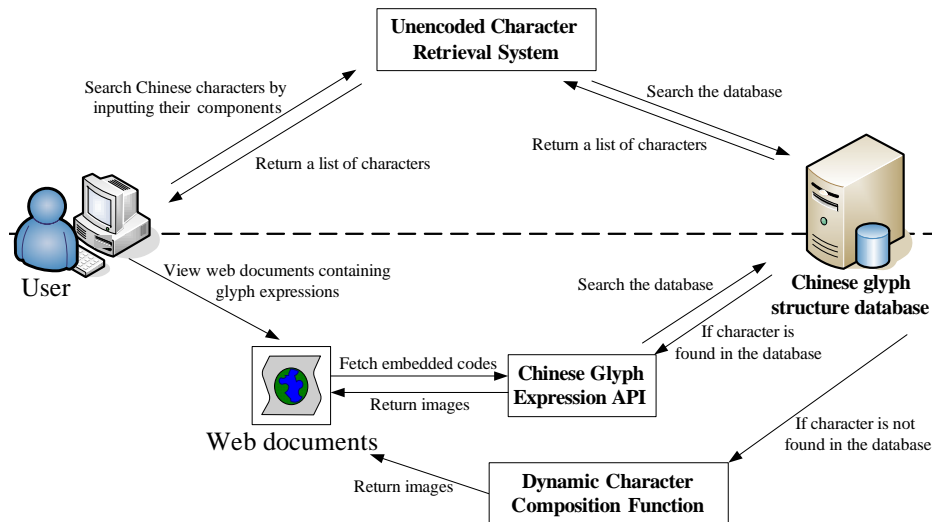


Fig. 1. An overview of the proposed mechanism

1. **Unencoded Character Retrieval System.** The system's function is to search and retrieve Chinese characters and related information from the database. Users can input the components that comprise a character, and the system will return a list of characters that contain the specified components, together with their corresponding glyph expressions and images of the characters. For example, if a user wants to find "𠄎" (glyph expression: "𠄎方𠄎土"), he can input "方" and "土" for the search.
2. **Chinese Glyph Expression API.** Web document editors can insert a glyph expression, which is a piece of Java Script code, into a web document to represent an unencoded character. When the web document is opened, the Chinese glyph expression API will fetch the glyph expression and send it to the database. If the database contains a corresponding character for the glyph expression it will return the character image to replace the glyph expression in the original web document. The image's background is set as transparent, and its size is dependent on the current browser's font size.
3. **Dynamic Character Composition Function.** If the database does not contain a corresponding character for a glyph expression, it will call the dynamic character composition function to generate a new character image immediately. For example, take the glyph expression "火𠄎中𠄎天," which does not exist in the database. In the case, the dynamic character composition function, which is available as open source software [5], will generate a character image "𠄎".

The following scenario illustrates the use of the proposed mechanism. If users encounter the unencoded character problem when editing a web document, they can search the database to find the appropriate glyph expression and an image of the input unencoded character. The glyph expression or the image can then be embedded in the

web document. If other users browse the document, the glyph expressions will be processed by the Chinese glyph expression API, which transforms them into character images, even though corresponding characters exist in the database. As a result, users can view images on different platforms without installing additional font files in their personal computers.

The proposed web-based mechanism, which is now available online, has been adopted by the National Digital Archive Program (NDAP), Taiwan [4]. Currently we serve many institutions and users in NDAP, such as Academia Sinica, the National Palace Museum, and the Ministry of Education. The following statistics shows the number of times the web-based service was accessed by specific NDAP projects up to January, 2008: "Name Authority File Project" - 1,111 times; "The Ancient Book Project" - 777,655 times; and "The Utensil Project" - 44,027 times.

3 Future Work

Although the primary function of the proposed mechanism is to solve the problem of accessing unencoded characters in web documents, it can be integrated with various web applications. It can also be applied to other web programming languages. The mechanism has already been adopted by many users in NDAP. In the future, we plan to develop a website to provide knowledge on ancient and modern Chinese characters based on the current framework. Through the website, the content of the Chinese glyph structure database will be available on-line so that users can easily access Chinese character resources. The website will connect with online dictionaries and serve as a Chinese character search portal.

Acknowledgements: This research was supported in part by Taiwan E-learning and Digital Archives Programs (TELDAP) sponsored by the National Science Council of Taiwan under NSC Grants: NSC 96-3113-H-001-010, NSC 96-3113-H-001-011 and NSC 96-3113-H-001-012

References

1. C. C. Hsieh, "The Missing Character Problem in Electronic Ancient Texts (電子古籍中的缺字問題)". In the First Conference on Chinese Etymology, Tianjin, Aug. 25-30, 1996. (in Chinese)
http://www.sinica.edu.tw/~cdp/paper/1996/19960825_1.htm
2. D. M. Juang, J. H. Wang, C. Y. Lai, C. C. Hsieh, L. F. Chien, and J. M. Ho, "Resolving the Unencoded Character Problem for Chinese Digital Libraries", Joint Conference on Digital Libraries (JCDL), p311-p319, Denver, Colorado, USA, June 7-11, 2005
3. Home of Chinese Document Processing Lab, <http://www.sinica.edu.tw/~cdp/service/>
4. National Digital Archives Program, <http://www.ndap.org.tw/>
5. Ksana Search Forge, <http://www.ksana.tw/>
6. Unencoded Chinese Characters System, <http://char.ndap.org.tw/>