Pacific Neighborhood Consortium (PNC) 15 May 1998

HANZIKU

An implementation of more than 54,000 Chinese Characters on mainstream operating platforms with compatibility of BIG5, JIS ,GB 2312 , KSC and Unicode encoded document.

漢字庫

實作五萬四千個中文字形於主流作業環境，並與通用編碼 (BIG5,JIS,GB 2312,KSC,Unicode)**文件相容。**

Ven. Huiray
Department of Research and Development, Fokuangshan.

Objectives of Hanziku:

1) To provide a substantial foundation for the maintenance of a Han-character database.
2) To provide an economical way for conversion between various contemporary Han-character mapping systems.
3) To provide a dictionary-position reference for computer defined Han characters, thereby establishing their forms, definitions, and pronunciations upon a literal source, so that variations in representation can be minimized.
4) To provide a solution for representation of Han texts across the board, avoiding the incompatibility of some conventional Han-character supplement systems such as those developed particularly for Buddhist canonical texts.

Areas of Collection:

1) Hanyu Dazidian
2) Buddhist canonical character supplements (Fokuang Chinese Buddhist Dictionary, Taisho and etc.)
3) All Characters found in Big 5, JIS, GB 2312, Unicode CJK ideography standard.
4) Character-parts developed by Academica Sinica, for the temporary expressing of uncollected characters.
5) Key-characters of various contemporary Chinese input methods. (eg. Changjei and Dayi)
6) Hangul, Hirakana, Katagana, Kanji, and Chinese dialectic characters.

Remarks:

1. The Han-character has been termed as the "Square-block", implying its inclusion of characters actually used by the Han race, as well as other square-blocks developed under its cultural influence. On this ground, the square-blocks of Hangul, Hirakana and Katagana are included in Hanziku.
2. Devanagari, European alphabets, and Diacritical mark are NOT included in Hanziku.
3. Although the styles of Han-characters are referenced to primeval glyphs, they are somewhat established upon the standard style, i.e. Kai (楷書), formulated after the restyling of Li.

Encoding Scheme:

Principle of Encoding:

1) **Minimum knowledge base is required** : A hard-copy of the dictionary is adequate for maintaining an undefined characters database, thereby reducing the overheads of its management.
2) **Use Font-planes to obtain greatest system compatibility** : The most popular contemporary Chinese operating systems are designed with not affecting the 13053 Big5 slots as a principle. To ensure smooth working under these operating systems, Hanziku limits itself only to use these 13053 slots. Currently there are 12 font-planes * 13053 slots = capacity of 156336 characters.
3) **Use virtual pages and loose encoding** : Corresponding to the provision of a dictionary-position code, Hanziku provides virtual pages of 32 characters per page to allow future insertion of additional characters according to their properties into appropriate sections (i.e. same radicals and strokes). Currently there are 156336 characters / 32 per page = 4894 pages. Characters are mapped to the Hanyu Daizidian within page 1 to page 4810. Pages 4811 to 4894 are the reservation for symbols and control codes.

Formats of Codes:

Each character having 3 parallel codes: a Hanziku-serial code, a Dictionary-position code, and a Plane-hex code. They can be considered as "Trinity", which are three different forms representing the same code or position.

1) The Hanziku serial code is an integer, e.g., 1 or 156636.
2) The Dictionary-position code is a 3-byte concatenation of a 4-digit page number, a period, and a 2-digit serial number, e.g., 1234.12

3) The Plane-hex code is a hyphenated concatenation of a font-plane number (1 to 9, and A-C) and a hex code, e.g., for Big5 , we have 1-A140 or C-F9D5

Conversion of codes :

1) Hanziku serial code = dictionary page * 32 + dictionary serial.
   e.g., Dictionary position code 1234.12 = Serial code 39500
   (Dictionary page 1234 * 32 + Dictionary serial 12 = Hanziku serial code 39500

   2)Font-plane = Integer portion of ((Serial code + Available slots) / Available slots) + 1
   e.g., Hanziku serial 1 = Font-plane 1, or Hanziku serial 156336 = Font-plane C
   ((Hanziku 1 + Available slots 13053) / Available slots 13053 = Integer 1)

((Hanziku 156336 + Available slots 13053) / Available slots 13053=Integer12=C)

3) Hex value = remains of ((Serial code + Available slots) / Available slots) + 1

e.g: Serial code 1 maps to Big5 A440, 13053 maps to F9D5, and so on.

Consideration of Compatibility with other Contemporary Encoding Systems:

1) Only utilizing regular code slots, such as the Big5' s 13053, and not the symbols or the user-fonts sections, thereby avoiding any mapping manipulation by the operating system, and ensuring the correct representation of Hanziku defined characters.
2) Adopting standard structures for font-files, such as TTF and Postscript.
3) There are two recommended ways to store the codes. One way is by embedding into the document marked dictionary-position codes, such as <1234.12>. This manner provides convenience when crossing platforms or shifting amongst various encoding systems. The other way is to use a 4-byte placeholder for each character, such as "!1 亡" can be used as a placeholder for the character "㇀" because the hex code of the character "亡" would generate the character "㇀" once it is mapped by Hanziku. This manner easily allows display of intended characters in editors through macro operations, such as converting "!1" into font-name specification, without having to engage extra programmatic efforts. There are pros and cons in both ways, but they can be adopted according to circumstances.

Consideration of Compatibility with Contemporary Applications:

1) The displaying of characters is achieved through font-name specification.
2) The current methods of inputting characters are maintained, and supplemented by a Hanziku character identification program.
3) Subsets of characters can be generated from the Hanziku master set. These subsets can bear Hanziku mappings, or user-font specific mappings if they are accommodated into the user-font sections. Regardless, the mappings are interchangeable.