

Title: Shuowen Seal Encoding Design Issues
Type: individual contribution
Source: suzuki toshiya, Richard Cook
Date: 2017-06-18

Preparing for the ad-hoc meeting on Shuowen (SW) Seal encoding, in this document we present a few questions and discussion points, for the consideration of experts contributing to the current draft of a proposal to encode Seal characters in UCS. We hope feedback will be available before the meeting, to help determine the meeting agenda.

Q1: Is the encoding of Seal as a separate script the best solution?

When Old Hanzi encoding was discussed and considered as another script in 2003, the experts were still unsure whether the variation selector is widely used, and how many devices implement it. Considering that Variation Sequences (VS) are widely supported, is UCS encoding (by means of independent code points) the best solution? Or, might a solution involving Standardized Variation Sequences (SVS) be better? If one solution is better, which and why?

One problem of course relates to the fact the UCS currently sometimes encodes the same abstract character multiple times. For example, the two code points 說[U+8AAA] 說[U+8AAC] both represent the same abstract character, and a third code point 說[U+8BF4] also represents this same abstract character (despite glyph differences). One of these code points might be selected as the Base for a VS to specify the Seal glyph. And the other two code points would map to the standard Base?

Q2: How should duplicate glyphs be handled in a standard encoding?





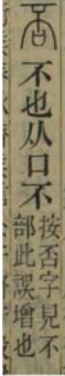
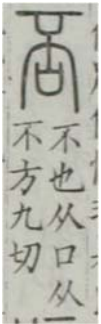


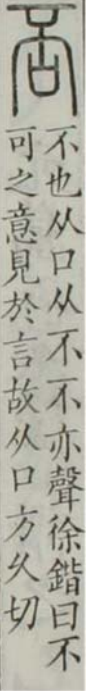
As mentioned in the past WG2 documents, many SW editions include several duplicate or possibly-duplicated characters in different positions.

For example, the character “否” occurs twice in many SW editions, once under 口 radical, and again under 不 radical. The radical assignments differ, but there is no semantic difference. If the same character is encoded twice, this will complicate search and collation. Thus, it is almost impossible for the users to choose which is appropriate for their contexts, except for the case in which one is typesetting 說文解字 as it is. But, the latest proposal of Small Seal encoding does not address this issue.

If duplicate characters are to be encoded, what will be the standard means for addressing search and collation issues? Note that for Seal (as for CJK) the definition of “duplicate” depends on several things, including especially the specific edition of Shuowen (from which the glyph is drawn), glyph style, and the associated property data for that edition (including radical assignment, gloss, commentary opinion).

Similar cases are found in the obsolete “Source Separation” guideline (Annex S) for CJK Unified Ideographs, and CJK Compatibility Ideographs. Today, both approaches are no longer good choices; source separation is no longer applied, the introduction of new compatibility ideograph is no longer welcome.

It might be more appropriate to consider SVS, or registration in IVD (Ideographic Variation Database).

口部(卷 02)								不部(卷 12)							
N4688 (Appendix Volume2)								N4688 (Appendix Volume12)							
255	01005			否	口	22	Zhengzhuan	6	08352			否	不	432	Zhengzhuan
段注本		藤花樹本						段注本		藤花樹本					
															

Duplicated “否” issue

Q3: Compatibility with existing reference and research implementations.

In our observation, 陳昌治本 would be the most widely referred version of 大徐本; for example, in Japanese academic libraries, the reprint of 陳昌治本 from 中華書局 are found in 103 universities (1963 ed), plus 9 universities (1972 ed). The numbers of the libraries holding 藤花樹本 and its reprint (by 商務印書館) is less than 10. Also, checking the calligraphic texts for Seal script in Japan, most of them refer 陳昌治本 (when some versions are specified). None of them refer 藤花樹本. If the situation is different in China mainland and Taiwan, please show with the evidence.

The situation described by a statement such as "all 藤花樹本 glyphs are coded, but we don't know which 陳昌治本 glyphs could be represented by 藤花樹本 codepoints, which glyphs could be separately coded in future" is not good situation.

Even if current working experts cannot afford to provide the fonts for 陳昌治本, which glyphs should be coded by 藤花樹本 codepoints should be clarified, at least. If Chinese experts have a consensus as it is not essential, it should be rationalized with the comments from some authorized society of Old Hanzi scholars.