

Title: Response to WG2 N4992

Source: Selena Wei

Date: 2018-06-20

Type: individual contribution

On the three issues brought up by Suzuki San, my understanding and suggestions are as follows:

The first is the encoding of Guwen (古文) and Zhouwen (籀文) in *Shuowen*, according to *Shuowen preface*, Xu Shen 許慎 mentioned that his principle of collecting characters is: "mainly collect Seal scripts, but Guwen (古文) and Zhouwen (籀文) are also accepted. (今敘篆文，合以古籀)" (Vol. 15) Therefore, Small Seal, Guwen (古文) and Zhouwen (籀文) were all included in Xu Shen's 許慎 collections; While Small Seal is treated as Zhengzhuan (正篆), the others are classified in the "Variant" (重文) section. Considering that *Shuowen* is not only a dictionary but also a comprehensive sorting and build-up of the small seal glyph system; therefore, even if glyphs that have only subtle differences are worth keeping, and they should be encoded individually. No matter how many kinds of scripts there are in *Shuowen* (說文) that can be academically considered different, we have always regarded them as an inseparable unified whole. Encoding all characters of *Shuowen* (說文) in the same block would be more convenient for the computer to make encoding judgement while processing the relevant content.

It is an issue in the academic circles to determine whether a character is included in *Shuowen*, but it is not necessary for the users to deal with this issue. Uniformly encoding *Shuowen* characters in a block requires only one simple space in the computer. If they are divided into several hundred blocks, and then add to different sources, the code points between each block cannot be correspondent as uniformly as those in Southeast Asia characters and South Asia characters. Then, in order to judge

whether a character is included in *Shuowen* can only be solved by giving additional conditional retrieve. Although current technology is capable of doing this, there is really no need to do things in a difficult way while it can be done in an easy way.

The second is the repertoire: Adding other sources also means taking longer time to develop. *Shuowen* is a complete system, adding new sources is just like the same matter of doing things in a difficult way instead of making it easy. Moreover, *Shuowen* has been circulated up to now, and many scholars have scrambled to study it, the content of the book can also be said to be an important category of the modern linguistics. Therefore, the probability of using the characters of *Shuowen* is undoubtedly much higher in the past. Even for the Small Seal scripts, there are more included in *Shuowen*.

The third is normalization: Suzuki San mentioned the NFKD/NFKC normalization issue in his document. I agree that it is a point and we could be discussing in future. NFKD/NFKC is not to be discussed at the present stage. At this stage, encoding should be handled first. According to my personal opinion, I do not agree to use NFKD because CJK is not using NFKD.

Finally, WG2 N4973 was jointly completed by the Small Seal scholars and experts from China and TCA. It is sufficient to represent the opinion of the academic community. It has nothing to do with attending the meeting. We welcome suggestions from everyone and the opportunity to make discussion together.

(end of document)