ISO/IEC JTC1/SC2/WG2 N4716

Date: 2016/04/28

Title: Proposal to Apply Source-Based Variation Selector in Shuowen Small Seal Encoding

Source: SUZUKI Toshiya

Document Type: Individual Contribution

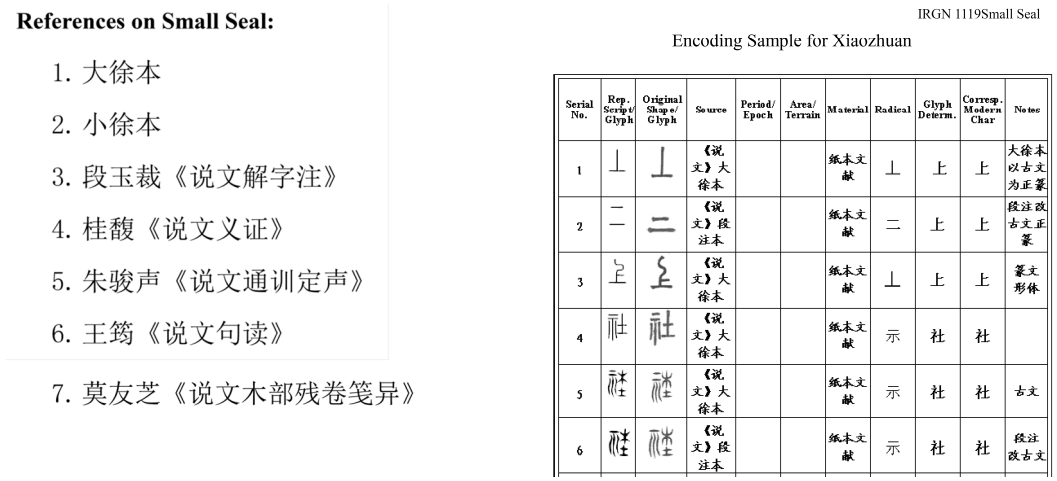
Distribution: WG2 Experts

Abstract

I propose an application of source-based variation selector in Small Seal encoding proposed by China NB and TCA (WG2 N4688). Because the proposal is not based on excavated materials, but almost entirely on the single book, Shuowen Jiezi, the shape-based separate encoding of the glyphs in each version would hamper information interchange.

1. Overview of Proposals of Small Seal Encoding

In the earliest proposal of Old Hanzi encoding, the project seemed to be designed to collect everything related with Small Seal and create a unified registry. The titles of the references listed in IRG N1134 are not only Shuowen itself but also selected authorized commentaries like 説文義証, 説文通訓定聲, 説文句読 and 説文木部残巻箋異.



earliest reference in IRG N1134 idea of unified glyph registry in IRG N1119

Figure 1: Earliest Design of Small Seal Encoding

In the totally revised proposal from 2014 (WG2 N4634), the source materials are reduced to 3 books; Shuowen Jiezi; Tenghuaxie version (藤花榭本, abbreviated as THX

in following), Chen Changzhi version (陳昌治本, 一篆一行本, abbreviated as CCZ in following) and Duan Yucai version (段注本, abbreviated as DYC in following). Pingjingguan version (平津館本, which is the source of CCZ, abbreviated as PJG in following) and a few XiaoXu versions (小徐本) were mentioned but excluded. In the latest proposal from 2015 (WG2 N4688), the sources have been reduced to only THX. Although the latest proposal does not clarify how the glyphs in other versions should be dealt with, it seems that TCA experts expect the separated encoding of the glyphs in other versions as long as their glyph shapes are different, without a variation selector mechanism, as described in WG2 N4634.

2. Examples of Glyphic Differences of Small Seal

Shuowen Jiezi is often quoted to design the official or orthodox Kaishu, Mingti and Sungti glyphs for printing typeface, so some people might be very sensitive for their structure. Yet there is no consensus how 2 glyphs could be identified as unifiable or not in WG2 nor IRG before, so here I list 3 types of glyphic differences of Small Seal in Shuowen Jiezi.

2.1. Differences Hard to Recognize (Type A)

Following glyph shape differences are subtle; crossing or touching differences. If the Small Seal calligraphers are learning the glyph shape with each stroke (as Kaishu), following glyphs could be recognized as differently written. However, these examples are taken from THX version, we could not find the consistency. Some people could expect they are differentiated intentionally, but there is no written explanation in the description, so it is very hard to memorize.

Figure 2: Variations of Small Seal "禹" in 藤花榭本 (THX)

2.2. Differences Easy to Recognize but Hard to Memorize (Type B)

Following glyph shape differences are legitimate; if somebody tries to express the structure by IDS, they should be different. However, again the difference is not per-version but per-character, and there is no explanation why some characters are in enclosing structure, others are in top-bottom structure. It would not be easy for the

users to memorize.

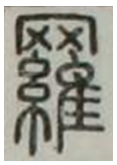
藤花榭本 (THX)					
陳昌治本 (CCZ)					

Figure 3: Structural Difference of "网"-Head Characters in 藤花榭本 and 陳昌治本

2.3. Differences Easy to Recognize and Memorize (Type C)

Following glyph shape difference are legitimate, and systematically designed. DYC version "improved" the shape of "門" (gate) from the conventional design in the earlier versions, because Shuowen Jiezi explains as the glyph to mean a gate consist of two doors (戶) facing each other. Therefore, all 門 in DYC versions are designed consistently. It is very easy for the experts to memorize how the glyph should be designed and when the glyph should be used. But, this situation is similar to the case "single dot 辶 and double dot 辶 should be distinguished?". Although DYC design is clearly differentiated, there is no semantic difference from the conventional glyph.

藤花榭本 (THX)				…
段注本 (DYC)				…

Figure 4: Systematic Modification of "門" Glyph in DYC version

3. Application of Variation Selector

3.1. Why Variation Selector? The Problems in Separated Encoding of Glyphic Variants

If the glyphic variations in above are coded separately, the representative glyphs should

be carefully chosen. The current proposal has chosen THX as the best reference to design the representative glyphs, because another candidate CCZ/PJG were claimed to be "revised", and inappropriate for the purpose "to retain the original contents as much as possible". But the rationale is insufficient, there are 2 points:

① If the version printed in Sung dynasty is the best, why real Sung dynasty version was not chosen? 續古逸叢書 and 四部叢書 have published the photocopied images of 青浦王昶本 (= 陸心源皕宋樓本 = 静嘉堂岩崎本), 中華再造善本 project have published the clone of 海源閣本 (currently preserved in Beijing National Library). I don't know which is better, but I believe they are widely available in China mainland and Taiwan.

② Is THX guaranteed as the most truthful copy of Sung dynasty version? I know some scholars had once said such, for example, Takada Jonosuke (高田襄之介: "中國字書史の研究" (ISBN 9784625420153, 1979, 明治書院, Japan, p.187) wrote as "although THX includes more errors than PJG, there is no errors introduced during the revision. This book is respected because it is rare than PJG". It is questionable whether it says the content of THX is better than PJG[1]. In addition, I could not find any concrete comparison by listing the differences from Sung dynasty version. Furthermore, the identification of the source material used by THX is currently controversial. When the scholars had the difficulty to access the materials, 海源閣本 was believed to be the source of THX, but recently there is a report that it should not be, because the lack of expected ownership stamp and the difference of the content (王貴元: "《説文解字》版本考述", 古籍整理研究学刊, 1999 年第 6 期, p.41-43, p.34). If his concern is reasonable, the evaluation "THX is the best version retaining Sung dynasty material" would be regarded as unfounded, because we don't have the source material of THX.

是汪中所藏宋小字本，书末有道光十八年(1838)丁晏跋文。原本后归山东聊城杨氏海源阁，杨绍和写有题识，谓"藤花榭所据之宋椠，即此本也"，今以二本对校，不同处特多。此本内有"额勒布号约斋"、"额勒布印"等印迹，知曾为额勒布收藏，但据额勒布藤花榭本序，藤花榭本所据为新安鲍惜分家藏宋本，而此本之内并无鲍惜分印迹，则鲍惜分未必收藏过此本。因此，藤花榭本所据宋本定非此本。今存第三种宋本缺标目，内有多枚"黄氏志淳"篆文朱文方印。

Comment by 王貴元 on the Relationship between 海源閣本 and THX.

Even if we have a consensus about the best one for its similarity to Sung dynasty

[1] 周祖謨 made a detailed comparison list between PJG and 青浦王昶本, but he evaluated PJG better than THX [13].

version, existing Small Seal fonts should not be disadvantaged. Once we decided THX glyphs are representative and different glyphs should be encoded separately, font vendors would be forced to search for incompatible glyphs in their existing products and decide whether to eliminate these glyphs, or propose to allocate new code points for these glyphs. It looks like "standardization does not help the implementation or migration, just help the obsoleting of the existing product", it is not good idea.

3.2. Proposal to Apply the Idea of Source-Based Variation Selector

If we consider the applicability of the variation selector mechanism, the glyphic variations in the type B & C would not be so difficult to make a registry listing the concrete glyphs. However, making a list for the glyphic variations for type A would not be easy, because there is no stable discussion about the unification for type A difference. In addition, the utilization of the shape-based variation selector for type B would not be easy. It is suspicious whether the users of coded Small Seal can know the detailed shape of the glyph, without seeing available options; the glyphs are not consistently designed, and no rationale of the design is given. To guide the users, the implementations would be requested to supply all registered variations, to show available options visibly. Its implementation cost would be expensive.

Recently Ken Lunde proposed the variation selectors for the regional customization without considering detailed glyph shape (L2/16-063), and it proposes a pseudo-regional variation selector to specify Kangxi-like shape (XK). It would be less-confusing to specify a version of Shuowen Jiezi other than by their shapes, and it would be useful to prevent the assignment of the variation sequence for mistakenly designed fonts.

So, I propose to apply source-based variation selectors for Small Seal script; the variation selectors should specify THX, PJG, CCZ, DYC, and more if additional versions are needed. Even if we consider several XiaoXu versions and commentaries which were considered in the earliest project, the number of versions to be registered would not be greater than 50. Considering that most materials share the same texts, the character identification by their description would be far more stable than identification by glyph shape.

3.3. Demonstration of Source-Based VS

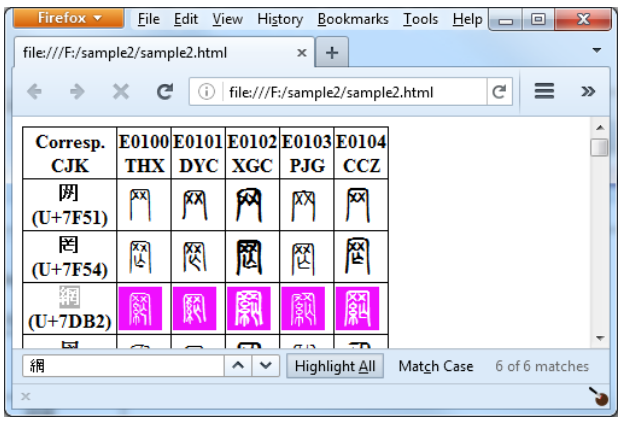
I built a sample font implementation that includes the glyphic variations of multiple versions of Shuowen Jiezi using variation sequences specified in the Format 14 ‘cmap’ subtable.

http://gyvern.ipc.hiroshima-u.ac.jp/~mpsuzuki/SWVS-20160427-2240.zip

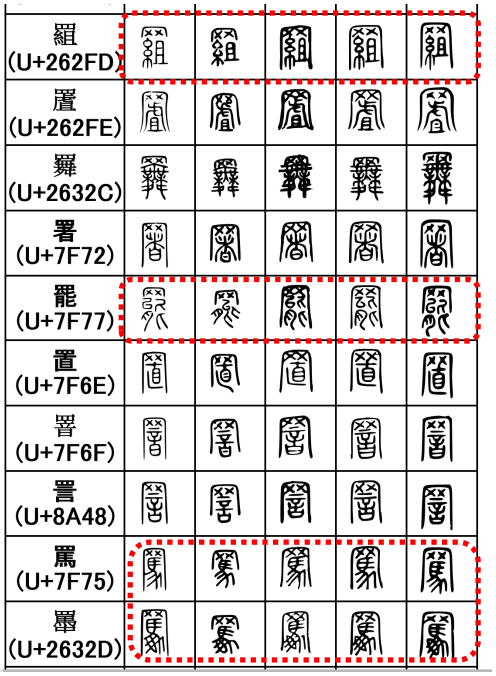
This zip file includes a sample font and a HTML file to show the glyphs in the sample font. Please extract and place both files in the same directory, then open the HTML file with the web browsers supporting web font and VS (I've checked the latest browsers; MS Edge, Firefox and Chromium). If your web browser is enabled to download the web font automatically, you can also try with;

http://gyvern.ipc.hiroshima-u.ac.jp/~mpsuzuki/swvs.html

It can search all variants by the base code point of the corresponding CJK Unified Ideograph, but displayed shapes are changed by the variation selectors. The code points in the sample font are taken from the corresponding CJK Unified Ideographs in WG2 N4688, and variation selectors (U+E0100 for THX, U+E0101 for DYC, U+E0102 for ZGC = 續古逸叢書, U+E0103 for PJG, U+E0104 for CCZ) are chosen only for temporary demonstration purposes. They are not intended as a proposal for assigning code points.



By searching the base character, all variations could be found.



per-version variants

Figure 5: Sample Implementation of VS with TrueType cmap format 14.

4. Other Issues: Duplicated Characters

Even if we identify a character by its description, instead of their glyph shapes, still there are several duplicated characters. For example, 右 (right) is once listed under the radical 口 (mouth), at volume 2 former part (卷 2 上). In later, 右 is listed again under the radical 又 (hand) at volume 3 latter part (卷 3 下). Current proposal is designed to encode them separately.

189	00939			右	口	22	Zhengzhuan
560	02078			右	又	76	Zhengzhuan

Figure 6: Duplicated Small Seal Characters in the Latest Proposal WG2 N4688

The description of 2 characters are quite similar. DYC tried to explain why similar characters are placed at 2 radicals, but it seems that he got no reasonable solution.

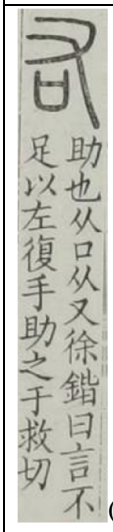
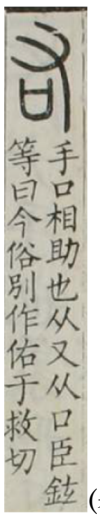
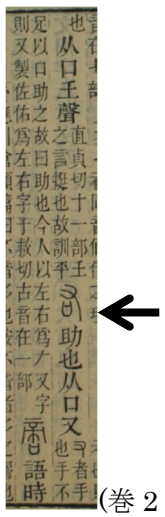
藤花榭本（THX）	段注本（DYC）
 (卷 2 上)  (卷 3 下)	 (卷 2 上)  (卷 3 下)

Figure 7: Descriptions of 2 右 Characters in HTX and DYC

Should we deal them as different characters? There are 2 difficulties to do that.

① The glyphs would be quite similar in most versions.

② The descriptions are quite similar (「助也从口又」「助也从又口」). If we are the machines comparing the text by strcmp(), we could say “their descriptions are different, so we could distinguish”. But we would not be able to choose an appropriate code point from two candidates for a given glyph, except of the case where we are simply recreating Shuowen.

The application of VS could not solve this issue easily. Further investigation on the existing electronic data (e.g. Old Hanzi indexed by Shuowen heading characters) is needed.

References

[1] China, "Old Hanzi Samples from PRC", 2005-05-18, ISO/IEC JTC1/SC2/WG2/IRG N1119

[2] China, "References on Old Hanzi", 2005-05-25, ISO/IEC JTC1/SC2/WG2/IRG N1134

[3] TCA and China, "Proposal to encode Small Seal Script in UCS", 2014-09-30, ISO/IEC JTC1/SC2/WG2 N4634

[4] TCA and China, "Proposal to encode Small Seal Script in UCS", 2015-10-20, ISO/IEC JTC1/SC2/WG2 N4688

[5] Ken Lunde, "Proposal to accept the submission to register the "PanCJKV" IVD collection", 2016-03-10, L2/06-163

[6] 仿北宋小字本説文解字　嘉慶丁卯年開雕　藤花榭藏板(藤花榭本)

[7] 新刻説文解字附説文通檢(陳昌治本)

[8] 経韻楼藏版　説文解字注　六書音均表附(段注本)

[9] 續古逸叢書　宋本説文解字(青浦王昶本 = 岩崎本)、涵芬樓

[10] 四部叢刊経部　説文解字(青浦王昶本 = 岩崎本)、商務印書館, 1922

[11] 中華再造善本唐宋編經部　説文解字(海源閣本 = 丁晏跋本), 北京図書館出版社, 2004, ISBN 7-5013-2262-7

[12] 高田襄之介: "中國字書史の研究" (ISBN 9784625420153, 1979, 明治書院

[13] 周祖謨: "説文解字之宋刻本 - 孫刻説文解字校勘後記",中華書局, 1966, ISBN 7101043518, 上巻, p.760-800

[14] 王貴元: "《説文解字》版本考述", 古籍整理研究学刊, 1999 年第 6 期, p.41-43, p.34

(end of document)