WG2 N5209
Date: 2023-03-11

Title: Considerations concerning the Small Seal encoding initiative

Source: Michel Suignard, Project Editor
Status: Individual Contribution
Distribution: WG2

Summary: This document examines the recent Small Seal documents submitted last year by TCA/China and Richard Cook. It then compares their content and proposes a way forward leading to the encoding of the Small Seal repertoire in ISO/IEC 10646 and Unicode in a future version of these standards.

Context

The context of Small Seal encoding has been explored for many years and there is not the intent in this document to provide another point of view on the technical merit of the various proposals. The reader can consult the following page at https://www.unicode.org/L2/topical/seal/ to find all relevant documents and is expected to be familiar with their content. The main intent of this document is to compare the last two contributions and explore their commonality and propose a solution to move forward.

Terms

Sources: Small Seal sources are typically grouped in three sets based on some famous authors of the Shuowen Jienzi repertoire. The following list is a much-simplified view that should be sufficient in the context of this document:

1. Xú Xuàn, book known as Daxu Ben, of which multiple versions are derived: Tenghuaxie version (THX), Pingjinguan, and Chen Changzhi (CCZ). The set has 11108 elements and is recognized as the 'X' source in this document.
2. Xú Kǎi, book known as Xiaxu Ben. The set has 10724 elements and is recognized as the 'K' source. One known edition is by Qi XiZao (1839), (QJZ).
3. Duàn Yùcái, book known as Duan Zhu. The set has 10706 elements and is recognized as the 'D' source (DYC).

These sources may have multiple versions and editions, but in general experts seem agreeable to present a single view for each source. All these sets also include 540 'radicals' or classifiers which are used to order/classify the sets but are themselves part of the overall sets. Sometimes the radical set is duplicated as a separate entity which may give the impression that a given Small Seal set is 540 larger than it really is. Unlike CJK Unified ideograph radicals, there is no consensus to encode them separately. They always appear as the first member (and sometimes only member) of their group.

The latest proposal from TCA/China was made in June 2022 (document n5191), while Richard Cook created a source mapping document in November 2022 (L2/22-279). When compared, these documents are referred to as TCA/China proposal and Cook proposal.

Status

In general, most of the work has been concentrated on the 'X' source, and more specifically on the THX version. The THX version has been used as the primary index with corrections and improvement based on the CCZ version.

Therefore, in principle it should be considered as ‘single source’. Some elements of the ‘D’ source and ‘K’ sources were also considered for glyph improvement (for example in document n5187) but without creating a separate source.

In June 2022, TCA and China proposed a set based on THX, but with 17 entries removed after unification for a total of 11091 characters (11108-17). Many glyph adjustments were also made in response notably to WG2 n5133. The main contribution is document WG2 n5191 which provides a table including indexes, representative glyph, modern character equivalent, and the radical (shown in modern form).

In November 2022, Richard Cook provided a source mapping document (L2/22-279) comparing the 3 sources mentioned above. It does not, however, provide any glyphic or graphemic evidence which is rather unfortunate. It is somewhat related to the TCA/China contribution, but some exploration is required to detect the commonality.

Analysis of the contributions

These contributions converge in many aspects. The X-THX source has the expected number of element (11108) in both sides, although it is indexed as 10706 THX proper elements and 402 extensions by TCA/China. Note the number of 10706 elements coincide with the size of the D source, but their content doesn’t match.

Considering the X source alone, one contribution (TCA/China) sees 17 variant pairs, while the other one (Cook) sees 33 variants pairs, but fortunately the definition of the 33 variants pairs fully incorporates the 17 variants pairs from the other document. However, in 6 cases out of the common 17 pairs, the preferred variant is reversed. At first approach the preferred variant chosen by TCA/China seems more optimal.

For example, considering the pair X914-X8016 which have values THX 881 and THX 7729 respectively in THX value:

7	07729	水	湩		tuo1/tuo4	河津也。在西河西。从水垂聲。	encode
	00881	口	湩		tuo4	口液也。从口垂聲。唾或从水。	

And considering that X914 (THX 881) is in the radical group #22 (first element is X867 (THX 834):

00867	00834	38362		口	口		22

And therefore encased with a group that uses that radical, but with none of the glyph appearance that relates it to that group (beside sharing the right component with previous):

00913	00880	38390		唾	口		22
00914	00881			唾	口	水	22
00915	00882	38391		唌	口		22

It seems to make more sense to remove X914/THX 881 as duplicate/variant and preserve X8016/THX 7729, not the other way around.

08015	07728	39F3F		溤	水		410
08016	07729	39F40		漷	水		410
08017	07730	39F41		瀍	水		410

In another case concerning the pair X537-X7268 (THX-519 and THX-7002)), these characters are considered variant by Cook but not by TCA/China:

00537	00519	38218		蠶	艸		12		X:537 X:7268 K:518 K:7013 D:515 D:6991
07268	07002	39C55		然;蠶	火	艸	382		X:537 X:7268 K:518 K:7013 D:515 D:6991

Because the Cook document does not have any glyph evidence, it is difficult to assess the situation and to make an educated decision.

The Cook document includes 3 sources (X, K, and D) with a total of 11163 entries, including 41 variant sets. Of these 41 variants sets, 28 involve X;K;D pairs, 1 X;K pair, 1 X;D pair, 3 X pairs, 6 K pairs, and 2 D pairs.

Because both documents recognize a common subset of 17 pairs, we can already eliminate 17 code points for a total of 11146. If we further eliminate the 8 pairs unique to K and D source, we could go down to 11138. Eventually, the addition of K and D sources would add 47 unique entries (i.e. not shared with X sources), or 55 if we don't agree with any of the unification exclusive to K and D sources.

This author created a new table merging the table provided along with the TCA/China contribution with the table provided by Cook. The first step was to flatten all the variant sets by adding all the variants entries suppressed by unification, the unification decision is carried out by a 'do not encode' flag) in a new column. These resulted in 11163 entries of which 25 (17+8) have the 'do not encode' flag set. As expected, that table is a superset of both contributions (TCA/China and Cook), adding 55 rows to the first contribution. (The Cook document mentions 11122 entries, which when considering the 41 variants sets which are only included in a single row in that document, correspond to the 11163 total number of entries (11122 = 11163-41).

As noted below, the size of the table may be increased in the future by adding the cases where disunification among sources could be justified (Cook document entries with qualifier flag indicating a major difference in appearance).

Some table fragments:

index		D index	K index	X index	THX index	Unicode	TTF Font	Corresp Modern Char.今字	SW Radical 說文部首	different radical 重文部首	SW Radical Number	Not encode	Variant
1	X:1 K:1 D:1	00001	00001	00001	00001	38000		一	一		1		
2	X:2 K:2 D:2	00002	00002	00002	00002	38001		一	一		1		
3	X:3 K:3 D-3	00003	00003	00003	00003	38002		元	一		1		
4	X:4 K:4 D:4	00004	00004	00004	00004	38003		天	一		1		
5	X:5 K:5 D:5	00005	00005	00005	00005	38004		丕	一		1		
6	X:6 K:6 D:6	00006	00006	00006	00006	38005		吏	一		1		
7	X:7 K:7 D+7	00007	00007	00007	00007	38006		上	丄		2		
8	X:8 K:8 D+8	00008	00008	00008	00008	38007		上	丄		2		
427	X:425 K:407 K:729 D:404	00404	00407	00425	00407	381A8		苗	艸		12		X:425 K:407 K:729 D:404
539	X:537 X:7268 K:518 K:7013 D	00515	00518	00537	00519	38218		蘿	艸		12		X:537 X:7268 K:518 K:7013 D:515 D:6991
746			00729						艸		12	Y	X:425 K:407 K:729 D:404
5235	X:5215 K:5108 D:5086	05086	05108	05215	05047	39453			宀		269		
7300		06991	07013	07268	07002	39C55		然:蘿	火	艸	382		X:537 X:7268 K:518 K:7013 D:515 D:6991

While all entries provided by TCA/China provide the radical and modern form (except for the entry corresponding X5215, THX 5047), none of the 55 entries without X entries which only exist in Cook document provides that

information. In most cases, the radical value can be determined by interpolation considering the radical values of previous characters. However, determining the modern representation needs to be done for all these 55 entries.

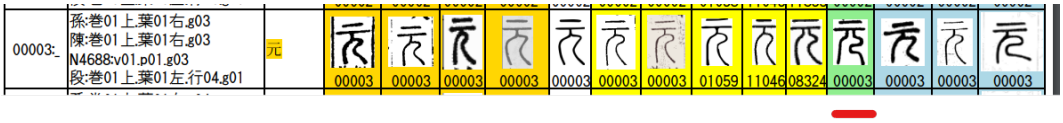
The table containing these 11163 is provided as link to this document in both pdf format (to see the actual glyphs) and in Excel format, check the reference section of this document for actual locations. The UCS code provided in the Unicode column are preliminary and are subject to change in revisions of this document. They were extracted from the TCA/China contribution.

Note that the mapping information from the Cook document also contains a qualifier flag indicating Major or Minor glyph differences between sources. This is marked by a '+' or '-' along the source values. For example, in the table above, the 7th entry has D-3, D+7, and D+8. Among these, some of the '+' markings could lead to further dis-unifications among the X, K, D sources.

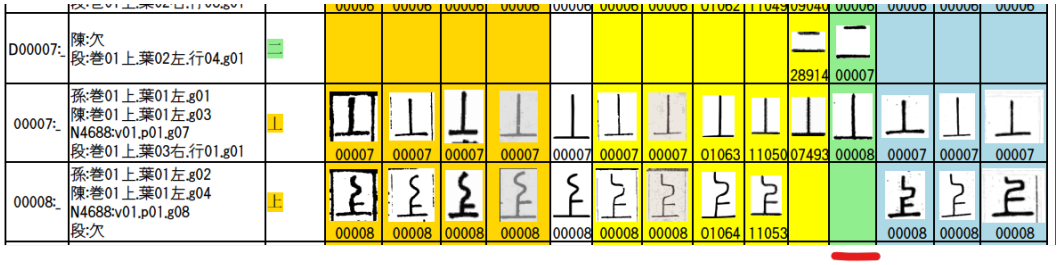
Finally, a great help is available through a massive file prepared by Toshiya Suzuki that describes graphically various editions of the 3 major sources with 14 columns, these shows graphic details that help understanding the qualifier flags present in the Cook document.

Examples related to the previous table:

Minor difference for the D source: X:3 K:3 D-3 (Orange-yellow: X sources, Green: D sources, Blue: K sources)



Major differences for the D sources: X:7 K:7 D+7 and X:8 K:8 D+8:

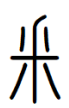


There are around 500 entries among the 11122 entries of the Cook document that have such flags so it only affects a minority of the entries. Therefore, the issue of further disunification can be simply addressed by only studying these 500 or so entries.

That specific case is about the dual representation of the radical #2 is interesting on its own as the typical representation as a regular Shuowen character is ⊥ while it tends to be represented as 二 when shown in a radical table, that form is a glyphic variant present in the D source along with the other form. Whether the two forms should be encoded is an open question.

Next step

The next step was to consider the collected data, considering for example the following fragment covering X799 to X806 (or THX 768 to THX 775):

D index	K index	X index	THX index	Unicode	TTF Font	Corresp Modern Char.今字	SW Radical 說文部首	different radical 重文部首	SW Radical Number
00763	00774	00799	00768	3831E		余	八		16
00764	00775	00800	00769	3831F		釆	釆		17
00765	00776	00801	00770	38320		釆	釆	丿	17
00766	00777	00802	00771	38321		番;蹯	釆		17
00767	00778	00803	00772	38322		蹯	釆	足	17
00768	00779	00804	00773	38323		蹯	釆	米	17
00769	00780	00805	00774	38324		審	釆		17
00770	00781	00806	00775	38325		審	釆		17

The data contains 3 sources (X, K, and D), a proposed code point, glyph, modern character(s), radical (modern form), alternate radical (also modern), and radical number (1 to 540). This led to the creation of the following records for each code point:

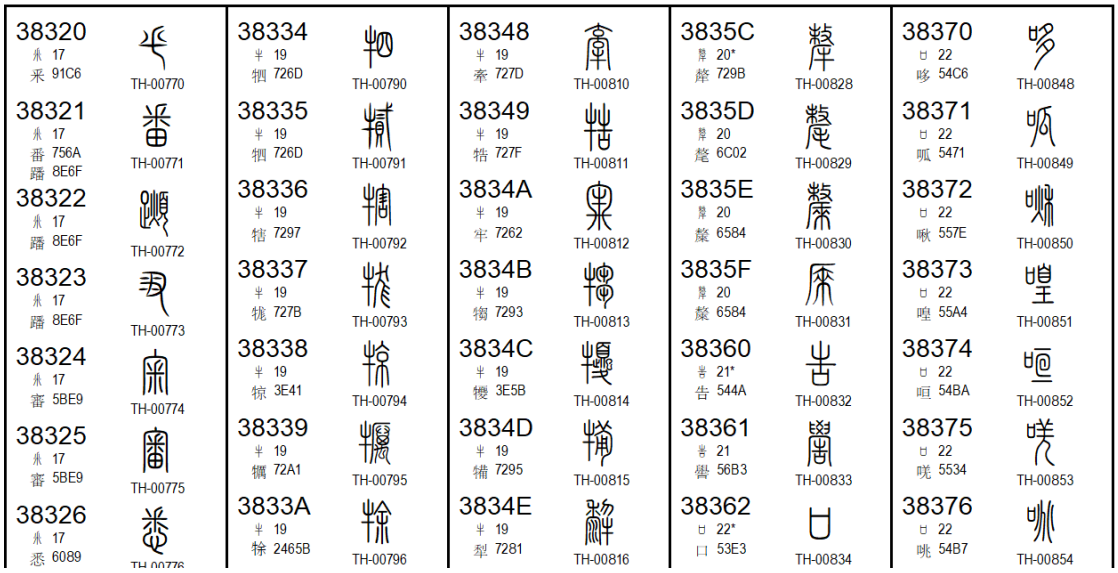
- kSEAL_THXSrc corresponding to the THX source takes values in the form TH-ddddd or X-ddd
- kSEAL_MCJK corresponding to the modern CJK equivalent in hexadecimal format, can be multiple, space separated.
- kSEAL_Rad radical made of the number followed by a dot and its encoded value. As such a radical entry is detected by the fact that its code point is the same as its radical value. It is also the first member of its group.

Other sources could be added, to represent the K source (XiaoXu Ben) currently denoted as kSEAL_QJZ and D source (Duan Zhu) currently denoted as kSEAL_DYC; these format names may change. The table shown above results in the following data fragment (the block starts at U+38000):

```
U+3831E kSEAL_THXSrc   TH-00768
U+3831E kSEAL_MCJK     4F59
U+3831E kSEAL_Rad      16.38312
U+3831F kSEAL_THXSrc   TH-00769
U+3831F kSEAL_MCJK     91C6
U+3831F kSEAL_Rad      17.3831F
U+38320 kSEAL_THXSrc   TH-00770
U+38320 kSEAL_MCJK     91C6
U+38320 kSEAL_Rad      17.3831F
U+38321 kSEAL_THXSrc   TH-00771
U+38321 kSEAL_MCJK     756A 8E6F
U+38321 kSEAL_Rad      17.3831F
U+38322 kSEAL_THXSrc   TH-00772
U+38322 kSEAL_MCJK     8E6F
U+38322 kSEAL_Rad      17.3831F
U+38323 kSEAL_THXSrc   TH-00773
U+38323 kSEAL_MCJK     8E6F
U+38323 kSEAL_Rad      17.3831F
U+38324 kSEAL_THXSrc   TH-00774
U+38324 kSEAL_MCJK     5BE9
U+38324 kSEAL_Rad      17.3831F
U+38325 kSEAL_THXSrc   TH-00775
U+38325 kSEAL_MCJK     5BE9
U+38325 kSEAL_Rad      17.3831F
```

Which is partially visible in the following code chart fragment:

38320 **Small Seal** **38383**



Note that the radicals are shown in their traditional Small Seal shape, not the modern character which can be seen as the modern CJK equivalent attached to the radical entries. In the code chart above, radicals for #17, #19, #20, #21, and #22 can be seen, with the radical entries for #20, #21, and #22 can be found respectively at U+3835C, U+38360, and U+38362. They are denoted by a '*' following the radical number in the code chart (not part of the data because it can be automatically generated).

While the current code chart only shows the X source, the software can easily be updated to show the 3 sources. However in the absence of fonts to display these sources, there is little value to do so at present.

The most critical part at this moment is to get access to the 55 entries from the Cook document that do not exist in the X sources, not just in term of representative glyph, but also in term of the ancillary data such as radical, modern representation(s) and any alternate radicals. This would also allow other experts to validate the 8 unifications done in that group of 55 characters.

Ideally, fonts should be provided for the totality of the K and D sources. But it may be acceptable in a first version to only provide glyph representation for these 55 unique glyphs, leaving the other entries blank or with glyphs similar to the equivalent X source entries.

In the absence of delivery of these 55 glyphs in a reasonable time, it is the opinion of this author that the proposal made by TCA/China should proceed in the format suggested above.

References:

Latest proposal by TCA/China:

N5187: TCA and China Feedback on WG2 N5133, https://www.unicode.org/wg2/docs/n5187-TCAandChina%20Feedback%20on%20WG2%20N5133.pdf

N5188: About the future extension of other versions of Shuowen Seal, https://www.unicode.org/wg2/docs/n5188-Extension-of-other-versions-of-Shuowen-Seal%20.pdf

N5189: THX Seal glyph Correction Principles, https://www.unicode.org/wg2/docs/n5189-THX-ModificationPrinciples.pdf

N5190: THX correction glyph summary table, https://www.unicode.org/wg2/docs/n5190-THX-correction-glyph-summary-table.pdf

N5191: THX Shuowen Properties Table 藤本《說文》小篆字表 pdf: [14.5 MB]: https://www.unicode.org/wg2/docs/n5191-THX-Properties-Table.pdf ; Excel: https://www.unicode.org/wg2/docs/n5191-THX-Properties-Table.xlsx

Richard Cook:

L2/22-279, UCS Seal Script Source Mapping Data, https://www.unicode.org/L2/L2022/22279-ucs-seal-map.pdf

Toshiya SUZUKI

Multi-sources table with glyphs (408MB!) http://gyvern.ipc.hiroshima-u.ac.jp/~mpsuzuki/ShuoWen/public-20171206-2230.pdf

Data set related to this document

Table in PDF attachment format: https://www.unicode.org/wg2/docs/n5209-SealDBnew-20230307.pdf
Table in excel format: https://www.unicode.org/wg2/docs/n5209-SealDBnew-20230307.xlsx

---