Title: Considerations concerning the Small Seal encoding initiative

Source: Michel Suignard, Project Editor
Status: Individual Contribution
Distribution: WG2
Replace: WG2 N5209

Summary: This document builds on previous contributions concerning the Small Seal script submitted since the last WG2 meeting in June 2024. The main intent of this document is to present a repertoire and corresponding database based on the preceding contributions. It is the opinion of the author that the Small Seal script as presented in this document is now a mature repertoire ready for encoding, once a final review is done to confirm the disunification of a small set of variants. The repertoire considered for encoding contains 11,140 code points.

Context

The context of Small Seal encoding has been explored for many years and there is no intent in this document at providing another point of view on the technical merit of the various proposals. The reader can consult the Unicode [Small Seal Topical] and the latest document [N5293] to find all relevant documents and is expected to be familiar with their content. Most of the source information and the analysis is based on the source mapping document created by Richard Cook [L2/22-279] and the glyphic evidence presented in [N5211] by SUZUKI Toshiya and further refined by experts from China, TCA, Japan, and the USA.

Terms

Sources: Small Seal sources are typically grouped in three sets based on well-known authors of Shuowen Jienzi publications. 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 11,108 elements and is recognized as the 'THX' source in this document. In some occurrences, the THX and CCZ sources are differentiated because they use different index schemes and have subtle differences in their visual representation. The term 'X' source may also be used for generic referencing of either THX or CCZ, using the simple 'ddddd' numerical convention .
2. Xú Kǎi, book known as Xiaxu Ben. One well-known edition is by Qi JunZao (1839). The set has 10,724 elements and is recognized as the 'K' source (QJZ).
3. Duàn Yùcái, book known as Duan Zhu. The set has 10,706 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. 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.

Status

The previous consideration document by this author [N5209] documented the three sets corresponding to the sources mentioned above. The union of these 3 sets amounted to 11,163 entries, including 11,108 entries for the THX source and 55 other entries not part of THX. Just considering the THX entries, it was commonly agreed that 17 entries

were part of variant pairs and therefore could be unified. This resulted in a list of 11,091 candidate code points (11108-17) when only considering the THX source. These entries were described in [N5191]. The other 55 entries unique to the QJZ and DYC sources needed further study to determine which entries should be added to a comprehensive Small Seal repertoire. The document [N5273] determined that among these 55 entries, 8 were duplicates (variants), and 47 entries should be added to the merged list. Furthermore, two additional entries, resulting from disunification between the three sources, were added to the list. Consequently, 49 entries were added to the list of proposed code points for a new total of 11,140 code points (11091+49). Note that with these 2 disunifications, the total number of entries in the index list is now 11,165. That index list is available as a separate file available in both [N5294-PDF] and [N5294-XLS] Excel format.

Finally, the document [N5293] provided representative glyphs for the additional 49 entries, therefore proving full glyph coverage for the 11,140 code points. In addition, a new supplementary font was also provided for these new entries. All this made it possible to create a draft repertoire for the Small Seal script including the 3 sources.

Presentation of the repertoire

Based on these three sets, it makes sense to present the repertoire in three fixed columns layout given that the three sources all include most of the repertoire. At this moment, we only have a full font for the THX source. For the other two sources, we have a combined font representing the unique 49 glyphs. These glyphs which may be shared between the QJZ and DYC sources tend to favor the QJZ representation when both contain a source reference to the same code point.

Currently, the code chart uses the THX glyph representation for all three sources as appropriate. When no THX glyph exists (for all entries without THX source), the same glyph from the supplementary font is used for the two remaining columns if both have a source reference. The editor is seeking availability of fonts for these two source. While the absence of these glyph references for the two sources is unfortunate, showing these additional glyphs is useful because they relate to the additional source references, facilitating visual matching. It should be noted that the Unicode [Small Seal Topical] resource provides additional visual representations of these sources as well.

Example showing various multi-column representations:

In addition, the code chart shows the radical codes (1 to 540) with their visual representation in Small Seal glyph version and the modern CJK ideograph code point and glyph associated with a given Small Sea code point. Both radical and modern values can be multiple. The code points corresponding to radical entries are identified by having a '*' near their radical number and are always the first code point in their group. In the fragment above, U+388CF has two modern values and U+388E6 is the entry for the radical #91.

For access to the code chart, please check the pointer provided in the References section below [N5294-Chart].

Small Seal Database

For each proposed code point, the database contains source references for up to 3 sources (THX, QJZ, and DYC), one or two modern characters, and one or two radical numbers (1 to 540). This has led to the definition of the following records for each code point (one source minimum, all code points have at least one modern representation and one radical):

- kSEAL_THXSrc corresponding to the THX source takes values in the form TH-ddddd or X-ddd,
- kSEAL_DYCSrc corresponding to the DYC source takes values in the form D-ddddd
- kSEAL_QJZSrc corresponding to the DYC source takes values in the form K-ddddd
- 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 code point corresponding to the 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. In some rare cases, there may be two radical values separated by a space character.

The code chart representation above corresponds to the following data:

```
U+388CF kSEAL_DYCSrc    D-02163                 ...
U+388CF kSEAL_QJZSrc    K-02184                 U+388E2 kSEAL_THXSrc    TH-02196
U+388CF kSEAL_MCJK      23AA9 6BBA              U+388E2 kSEAL_DYCSrc    D-02182
U+388CF kSEAL_Rad       87.388CE               U+388E2 kSEAL_QJZSrc    K-02203
U+388D0 kSEAL_THXSrc    TH-02179               U+388E2 kSEAL_MCJK      76B0
U+388D0 kSEAL_DYCSrc    D-02164                U+388E2 kSEAL_Rad       90.388DF
U+388D0 kSEAL_QJZSrc    K-02185                U+388E3 kSEAL_THXSrc    TH-02197
U+388D0 kSEAL_MCJK      6BBA                   U+388E3 kSEAL_DYCSrc    D-02183
U+388D0 kSEAL_Rad       87.388CE               U+388E3 kSEAL_QJZSrc    K-02204
U+388D1 kSEAL_THXSrc    TH-02180               U+388E3 kSEAL_MCJK      76AF
U+388D1 kSEAL_DYCSrc    D-02165                U+388E3 kSEAL_Rad       90.388DF
U+388D1 kSEAL_QJZSrc    K-02186                U+388E4 kSEAL_THXSrc    TH-X080
U+388D1 kSEAL_MCJK      6BBA                   U+388E4 kSEAL_MCJK      76B8
U+388D1 kSEAL_Rad       87.388CE               U+388E4 kSEAL_Rad       90.388DF
U+388D2 kSEAL_THXSrc    TH-02181               U+388E5 kSEAL_THXSrc    TH-X081
U+388D2 kSEAL_DYCSrc    D-02166                U+388E5 kSEAL_MCJK      76B4
U+388D2 kSEAL_QJZSrc    K-02187                U+388E5 kSEAL_Rad       90.388DF
U+388D2 kSEAL_MCJK      6BBA                   U+388E6 kSEAL_THXSrc    TH-02198
U+388D2 kSEAL_Rad       87.388CE               U+388E6 kSEAL_DYCSrc    D-02184
U+388D3 kSEAL_DYCSrc    D-02167                U+388E6 kSEAL_QJZSrc    K-02205
U+388D3 kSEAL_QJZSrc    K-02188                U+388E6 kSEAL_MCJK      3F31
U+388D3 kSEAL_MCJK      20B81 6BBA             U+388E6 kSEAL_Rad       91.388E6
U+388D3 kSEAL_Rad       87.388CE
```

Additional data is typically available for each code point in the research material, such as the modern form of the radical and an alternate form of the radical in modern form. The modern form of the radical can be deducted from the existing database (it is simply the modern form of the radical entry) and therefore does not need to be added. However, in the research material, the alternate forms (common) are not synchronized with the existence of multiple radicals (rare) for a given code point. To avoid inconsistency, these alternate modern radicals are not part of the database. Ideally, they should all correspond to multiple radical entries. Finally, whether or not to add a CCZ source is open to debate because while useful, there is a simple mapping between the two sources and their visual representations are very similar.

For access to the database, please check the pointer provided in the References section below [N5294-TXT].

Variants

Originally, the number of variant sets identified by the Cook document [L2/22-279] was 41. However, one additional variant was added later [N5273] between the following two sets of sources:
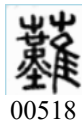
⼐ (source THX-6565 and K-06577) and ⼐ (source D-09753).

On that total of 42 variant sets, 17 were initially agreed by all experts, and 8 more were also agreed later (see document [WG2 N5273]). This leaves 17 variant sets which were deemed not unifiable by the latest contributions. These 17 sets are documented in the following pages (X and THX index represent the same source, the X index only uses a 5-digit notation for all sources and represents the CCZ entry, unlike CCZ the THX index uses a X-3-digit notation to show the extended repertoire).

To facilitate the review, the visual representation of the Small Seal sources documented in [N5211] are provided in these tables. For the D source, the column #11 of that document has been used as this is the only DYC column available. For the K source, column #14 has been used because its indexes match the indexes provided in [L2/22-279]. For the X sources, we use two column from [N5211], the column #3 represents the CCZ sources (CCZ index), and the column #5 represents the THX source (THX index). Because the THX sources are in fact equivalent to the current TTF font glyphs, they are not duplicated in the THX columns in the tables below.

Reviewers should confirm the status of these 17 variant sets. Any unification decision can be easily processed and result in an updated repertoire. Note that some sources may be moved between the code points as appropriate.

List of the 17 disunified variant sets

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:537 X:7268 K:518 K:7013 D:515 D:6991	00515	00518	00537	00519	3821A		蘿	艸		12
X:537 X:7268 K:518 K:7013 D:515 D:6991	06991	07013	07268	07002	39C72		然；蘿	火	艸	382

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:658 X:3641 K:639 K:3536 D+636 D:3514	00636	00639	00658	00640	38293		蒩	艸		12
X:658 X:3641 K:639 K:3536 D+636 D:3514	03514	03536	03641	03527	38E3C		蒕	血		173

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:659 X:3642 K:640 K:3537 D+637 D:3515	00637	00640	00659	00641	38294	蠶	菹	艸		12
X:659 X:3642 K:640 K:3537 D+637 D:3515	03515	03537	03642	03528	38E3D	蠶	蘁	血		173

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:965 X:6305 K:938 K:6097 D:927 D:6078	00927	00938	00965	00932	383C5	鬚	嘯	口	欠	22
X:965 X:6305 K:938 K:6097 D:927 D:6078	06078	06097	06305	06088	398AA	鬚	嘯	欠		320
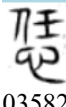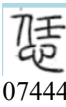
Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:1262 K:1224 K:8594 D-1212	01212	01224	01262	01218	384ED	赴	徙	辵		33
K:1224 K:8594		08594			3A2E4	徙	撠	手		441

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:1407 X:6235 K:1356 K:6029 D:1342 D:6009	01342	01356	01407	01350	3857E	㡾	得	彳	見	34
X:1407 X:6235 K:1356 K:6029 D:1342 D:6009	06009	06029	06235	06019	39865	㡾	尋	見		318

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:2068 X:3543 K:1999 K:3441 D:1977 D:3419	01977	01999	02068	01994	38812	鞈	鞈	革		70
X:2068 X:3543 K:1999 K:3441 D:1977 D:3419	03419	03441	03543	03432	38DDA	鞈	鼚	鼓	革	161

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
K:2288 K:9903		02288			38937	㤷	㲃；壞	攴		92
X:10254 K:2288 K:9903 D:9886	09886	09903	10254	09887	3A82A	㤷	壞	土	攴	480

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:2689 X:6635 K:2607 K:6418 D:2584 D:6396	02584	02607	02689	02599	38A80	羑	羑	羊		114
X:2689 X:6635 K:2607 K:6418 D:2584 D:6396	06396	06418	06635	06407	399F6	羑	誘	厶	羊	348

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:3687 X:7706 K:3582 K:7444 D:3560 D:7420	03560	03582	03687	03573	38E6A	佞	飪	食	心	180
X:3687 X:7706 K:3582 K:7444 D:3560 D:7420	07420	07444	07706	07432	39E29	悡	悡	心		408

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:3832 X:10208 K:3723 K:9857 D:3701 D:9840	03701	03723	03832	03714	38EFB	亶	亶	亶		189
X:3832 X:10208 K:3723 K:9857 D:3701 D:9840	09840	09857	10208	09841	3A7FB	亶	墉	土	高	480
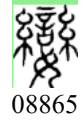
Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:3928 X:8935 K:3818 K:8615 D:3795 D:8601	03795	03818	03928	03809	38F5B		鞻	韋	手	201
X:3928 X:8935 K:3818 K:8615 D:3795 D:8601	08601	08615	08935	08611	3A2FA		挈	手		441

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:4208 X:10532 K:4089 K:10163 D:4072 D-10147	04072	04089	04208	04088	39074		椶	木		206
X:4208 X:10532 K:4089 K:10163 D:4072 D-10147	10147	10163	10532	10148	3A940		鏝；椶	金	木	490

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:5653 X:10981 K:5481 K:10602 D:5459 D:10585	05459	05481	05653	05470	3961C		保	人	子	287
X:5653 X:10981 K:5481 K:10602 D:5459 D:10585	10585	10602	10981	10585	3AB02		孟	子		525

Table 1

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:7934 X:8196 K:7686 K:7922 D:7898	07898	07686	07934	07647	39F0C		沇	水		410
X:7934 X:8196 K:7686 K:7922		07922	08196	07909	3A015		沿	水		410

Table 2

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:8347 X:10177 K:8069 K:9826 D:8049 D-9809	08049	08069	08347	08060	3A0AD		墢	水		410
X:8347 X:10177 K:8069 K:9826 D:8049 D-9809	09809	09826	10177	09810	3A7DC		墢	土		480

Table 3

Variants	D index	K index	CCZ index	THX index	Unicode	TTF Font	Modern	Radical	Other radical	SW Rad. Number
X:9210 X:9275 K:8888 K:8952 D:8865 D:8930	08865	08888	09210	08873	3A40F		變	女		443
X:9210 X:9275 K:8888 K:8952 D:8865 D:8930	08930	08952	09275	08938	3A450		變	女		443

Disunification

In document [N5273], the request was made to disunify the DYC (D) source in the two following entries because the DYC source glyphs had different structures than the other two sources. (The same document seems to have the modern representation for the index #664 in error, the table below uses the original values).

Index	D index	K index	CCZ index	THX index	TTF Font	Modern	Radical	Other radical	SW Rad. Number
664	00640	00643	00662	00644		蘫	艸		12
11071	10621	10638	11017	10621		申	申	日	536

This resulted in the following entries in the code chart

38297
艸 12
蘫 862B
TH-00644   K-00643

38298
艸 12
蘫 862B
D-00640

3AB26
甲 536
申 7533
TH-10621   K-10638

3AB27
甲 536
申 7533
D-10621

Index List

An index list containing all 11,165 raw entries is available for additional information. It contains mapping information initially provided by the Cook document [L2/22-279] and slightly updated to include missing entries and the two disunifications.

The table containing these 11,165 entries is provided as link to this document in both pdf format (to see the actual glyphs) and in Excel format (which will not contain the Small Seal glyphs), check the References section of this document for actual locations. The UCS (Unicode) code points provided in the UCS column correspond to the original values created in [N5191] and have been superseded by this document, as they include the newly proposed 49 code points. The main index range is keeping the original 1 to 11,163 value range to facilitate reference to current documents and the two additional entries are named 664A and 11071A.

The proposed Unicode code points in the code chart and the database are algorithmically generated in sequence by ignoring the rows where the 'Not Encode' column (not shown in the example below) is set to 'Y'.

Note that the mapping information from [L2/22-279] also contains a qualifier flag indicating Major or Minor glyph differences between sources which have been preserved in this new index table. This is marked by a '+' or '-' along the source values. For example, in the table below for index entries #3, #7, and #8, you can find D-3, D+7, and D+8. Among these, some of the '+' markings could lead to further disunifications among the X, K, D sources.

Index	Cook mapping	D index	K index	X index	THX index	UCS	TTF Font	Modern	Radical	Other radical	SW Rad. Number
3	X:3 K:3 D-3	00003	00003	00003	00003	38002	元	元	一		1
7	X:7 K:7 D+7	00007	00007	00007	00007	38006	丄	上	丄		2
8	X:8 K:8 D+8	00008	00008	00008	00008	38007	丄	上	丄		2

There are around 500 entries among the 11,122 entries of the Cook document that have such flags, therefore it only affects a minority of the entries. Accordingly, the issue of further disunification can be addressed by only studying these 500 or so entries.

References:

Overall reference content

Unicode [Small Seal Topical]: https://www.unicode.org/L2/topical/seal/ contains links to all related Small Seal documents in both Unicode and ISO/IEC WG2 registries up to document N5273.

Documents directly referenced in this document (may also be part of the topical content):

[L2/22-279]: UCS Seal Script Source Mapping Data, Richard Cook, 2022-11-08
https://www.unicode.org/L2/L2022/22279-ucs-seal-map.pdf

[N5191]: THX Shuowen Properties Table, TCA and China experts, 2022-0620 (**14.5 MB** pdf)
https://www.unicode.org/wg2/docs/n5191-THX-Properties-Table.pdf

[N5209]: Considerations concerning the Small Seal encoding initiative, Michel Suignard, 2023-03-07
https://www.unicode.org/wg2/docs/n5209-ConsiderationsSmallSeal.pdf

[N5211]: 14-Column Seal Script Glyph Comparison Chart – Documentation, SUZUKI Toshiya, 2023-04-03
https://www.unicode.org/wg2/docs/n5211-14-columnchart-documentation.pdf and related data (**420MB**)
https://www.unicode.org/wg2/docs/n5211-14-columnchart-20171206.pdf

[N5293]: TTF Font for 49 entries from K and D version, TCA and China experts, 2024-12-06
https://www.unicode.org/wg2/docs/n5293_TTF%20Font%20for%2049%20entries%20from%20K%20and%20D%20version.pdf

Appendixes (Data sets and Code Charts related to this document)

[N5294-PDF]: Raw Index List with 11,165 entries in PDF format (will show Small Seal Glyphs), 2025-01-17
https://www.unicode.org/wg2/docs/n5294-IndexPDF.pdf

[N5294-XLS]: Raw Index List with 11165 entries in XLSX format (it will not show the Small Seal glyphs), 2025-01-17
https://www.unicode.org/wg2/docs/n5294-indexXLS.xlsx

[N5294-TXT]: Small Seal database, 2025-01-17 https://www.unicode.org/wg2/docs/n5294-Database.txt

[N5294-Chart]: Small Seal Code chart, 2025-01-17 https://www.unicode.org/wg2/docs/n5294-SmallSeal-Chart.pdf

Acknowledgment