Title: UCS Seal Script Source Mapping Data
Source: Richard Cook <rscook@unicode.org>
Status: Individual Contribution
Action: For consideration by UTC and WG2/Seal Script Ad Hoc
Date: 2022-11-08T21:29:41Z

PREAMBLE:

This document and the associated mapping data (ucs_seal_map.txt) define conventions for the development of a Seal Script encoding model and a new block of UCS Seal Script characters. The mapping data here presented consolidates information on Seal Script assembled and proofed over the past 30+ years (32,538 Glyph IDs), into a standard framework for the design and production of UCS Seal Script code charts. WG2 Seal Script Ad Hoc (SSAH) group members have to-date compiled a variety of individual resources which as yet lack a common framework. The current document and data represent an effort to abstract away from single-source focus on the glyphs evident in one edition of one source (THX), to a more general perspective on multi-source mapping, quantifying the overall range of glyph variation, as is necessary to formulate an adequate abstract Seal Script character encoding model and repertory. Given the definitions and extensive mapping data provided here, SSAH contributors will be able to consolidate their glyph data for the production of draft multi-column UCS Seal Script code charts.

SOURCES:

The three primary source classes (X|K|D) of 《說文》 Shuōwén (SW) presented in the mapping data are here defined as follows:

X : 徐鉉 Xú Xuàn (916-991), a.k.a. 大徐 (elder of the brothers Xú); X is short for Xuàn (鉉);
K : 徐鍇 Xú Kǎi (920-974), a.k.a. 小徐 (younger of the brothers Xú); K is short for Kǎi (鍇);
D : 段玉裁 Duàn Yùcái (1735-1815); D is short for Duàn (段).

Each of these 3 source abbreviations (X|K|D) also refers to a primary bibliographic reference to a common modern print edition associated with each of the above authors:

X : 《說文解字·附檢字》（平裝本）○編撰：〔漢〕許慎；校定：〔宋〕徐鉉；出版 / 發行：中華書局（香港）有限公司 [1972.6; 1998.9; ISBN:

962-231-231-4; 1972.6, 1989.2, 1996.2; (精裝本) ISBN: 962-231-208-X].

K : 《說文解字·繫傳》(精裝本)○〔東漢〕許慎;〔南唐〕徐鍇選;北京:中華書局 [1987, 1998; ISBN: 7-101-00060-6/H.7].

D : 《說文解字·注》(精裝本)○〔東漢〕許慎;〔清〕段玉裁注;上海:上海古籍出版社 [1981; 1988, 1989 (9th printing); ISBN: 7-5325-0487 5/H.6].

In addition to the above-specified bibliographic sources (which have been imaged by various SSAH contributors, for glyph comparison), each source (X|K|D) also names a class of related editions to which the indexing conventions of the specified source apply. That is, sources in class X all use the exact same indexing scheme (SNs=GIDs, as defined below). Additional sources (XN|KN|DN) might be specified using explicit N (integer), perhaps to indicate indexing incompatibility. For example, the following editions are known to be compatible with the source class indexing (because the indexing of these electronic editions is complete), and so N might be omitted:

X1 : 仿北宋小字本《說文解字》藤花榭藏板○徐鉉撰○商務印書館摹印 [UCB EAL 5093.1914; THX, see <http://www.wul.waseda.ac.jp/kotenseki/html/ho04/ho04_00025/index.html>].

K1 : 《說文解字·繫傳》○徐鍇撰 [see <http://www.wul.waseda.ac.jp/kotenseki/html/bunko01/bunko01_01521_0070/index.html>].

D1 : 《說文解字·注》○段玉裁注 [see <http://www.wul.waseda.ac.jp/kotenseki/html/ho04/ho04_00026_0001/index.html>].

The following major texts have also been indexed for inclusion in future versions of project mapping data:

G: 《說文解字·義證》○撰:桂馥○濟南:齊魯書社出版發行 [1987, 1994; ISBN: 7-5333-0061-0/H.4; 1987, 1998; ISBN: 7-101-00065-7/H.12; see <http://www.wul.waseda.ac.jp/kotenseki/html/ho04/ho04_02178/index.html>].

DZ: 《說文大字典》○輯:沙青岩○刪編:汪仁壽、王鼎○天津:天津市美術出版社 [1980.7; UCB EAL PL1281.H83 S43; SWDZD.]

Z: 《說文解字·詁林及補遺》○編纂:丁福保○臺北:臺灣商務印書館 [1931, 1959, 1977, 1997; UCB EAL 5092.1032, PL1281.H83 T5 1977 RR; SWGL].

Note: G-source matches the X indexing with only minor exceptions; DZ is a hybrid edition widely used by modern scholars because it served as the basis for early BNU digitization efforts (it is a superset of X, sometimes substituting K forms (which sometimes are also D forms)); Z is a compilation reproducing many SW editions, with only general X indexing.

PROPERTY VALUES:

Property values for each source (S) are sequential decimal integer serial numbers (SN) also serving as Glyph ID (GID), derived from a sequential count (indexing) of all head entries in that source, with SN values per S in the following ranges (also associated with the following PUA code point ranges defined in the development implementation):

```
S : SN RANGE   : PUA CODE POINT RANGE
X : 1 .. 11108 : 0x102BEE .. 0x105751
K : 1 .. 10724 : 0x105752 .. 0x108135
D : 1 .. 10706 : 0x100000 .. 0x1029D1
T = 32,538 (GIDs)
```

Property values have a three-part form “S:SN”, in which: (1) “S” is a single value (X|K|D); (2) “:” is a delimiter (:|+|-); (3) SN is a numeric index value (GID) in the given range for S.

The default delimiter “:” of K and D sources may be replaced by a flag (+|-) indicating (+Major, -Minor) character/glyph differences (relative to the other sources); Minor differences are considered unifiable (and are not exhaustively flagged), but are sometimes marked to raise issues for general methodological discussion; Major differences (including differences in components and structure) are important to development of the encoding model and are exhaustively flagged, to serve as a guide to unification and disunification decisions.

FILE STRUCTURE:

The data file ucs_seal_map.txt has header and footer comment lines (beginning with “#”) including version and other information.

Each row of data in the file has 5 tab-delimited fields (0..4), with content as follows:

```
# 0  1    2    3    4
HEX  X    K    D    X:SN K:SN D:SN
```

0 : HEX : sequential hexadecimal row number, starting at 0x32400 (possible future UCS code point; see <https://www.unicode.org/roadmaps/tip/>);
1 : X : X-source PUA code point (UTF-8) for development (0x102BEE .. 0x105751);
2 : K : K-source PUA code point (UTF-8) for development (0x105752 .. 0x108135);
3 : D : D-source PUA code point (UTF-8) for development (0x100000 .. 0x1029D1);
4 : XKD : Space-delimited source mappings (S:SN); each source (X|K|D) may occur zero or more times (but at least one source must occur).

Field 4 matches the following regular expression:

```
/^[XKD][-\+:]\d+( [XKD][-\+:]\d+)+$/
```

Currently a total of 32538 source values (GIDs) are mapped, and the total number of data rows is 11122 (<= eventual candidate UCS characters).

The file footer comments include (1) the decimal integer value of the total number of data rows, and (2) EOF (End Of File) marking file termination.

REPERTORY:

The set of head entries in each source (X|K|D) is different and differently ordered, and the mappings between the sources are complex. Each entry is headed by a Seal form (篆文、籀文、古文) followed by an associated semantic gloss and structural analysis (pictographic forms may lack clear component analysis). This combination of Seal form and gloss/analysis constitutes a lexical mapping: the Seal form is associated with specific graph type, component structure and linguistic usage. In the textual transmission, the Seal forms sometimes vary, and the lexical mapping may also vary. For this reason the D-source sometimes emends the text in ways which complicate the inter-source mappings (and exemplify the kinds of changes the text has undergone in transmission over 2,000 years). Likewise, K-source underwent separate interpretation in transmission. The current

mapping data presents the set of lexical mappings between the three sources, flagging significant glyph differences to highlight lexical mapping issues, for discussion of encoding possibilities.

In addition to the set of lexical heads here mapped, several other Seal forms have been identified in the mapping process. For example, Seal-form components sometimes occur in-line in the structural analysis and lack head entries. An inventory of such forms will be provided in the future.

DUPLICATES:

Several duplicate forms occur within each source and have been mapped together in a single row. The same duplicate forms sometimes occur in other sources, and in such cases are also mapped together in the same row. In other cases duplicates in one source are not present in another. In the current mapping data, duplicates are to be found wherever a given source (X|K|D) is repeated in a given row of the mapping data. The inventory of duplicates was created after the indexing of each source, in conjunction with creation of the lexical and inter-source mappings. A typology of duplicate forms was also created, identifying the kind of duplicate (G=古文, Z=籀文, H=或体), true duplicates (vs. apparent duplicates or confusables), and duplicates which arose (or were eliminated) as a result of emendation. The lexical mapping is critical to determining the character identity, and so it also determines whether or not a pair of forms are true duplicates within a given source. Duplicates within a given source, but which are emended in another source eliminating that duplication, are candidates for code chart glyph correction (and complex lexical mappings), not simply elimination of the duplicate.

COLLATION:

Serialization of the lexical heads per source results in collation data for that source. That is, one may sort the entries occurring in that source back into the original source order simply by sorting the entries by serial number (SN).

Ordering of the rows in the current mapping data uses a fall-back method based primarily on the X-source. And where a form does not occur in the X-source, it is sorted after the previous form in that source which does occur in X-source. For duplicate forms, the first occurrence determines the sort position.

FUTURE:

With a standard framework for Seal Script mapping data, and given associated glyph data available from contributors, it is possible to generate multi-column code charts for working group discussion. In future phases of this work, with feedback from contributors, we will refine the analysis of glyph differences, extend the property data, and produce draft code charts, to define the encoding model. Future versions of this mapping data will be made available as resources permit.

EOF