ISO/IEC JTC1/SC2/WG2 N4844
2017-08-25

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**

Title: **Comments on encoding Shuowen Small Seal**

Source: **Selena Wei, Eiso Chan**

Status: **Individual Contribution to WG2 #66 and Small Seal Script Ad hoc Meeting**

Action: **For consideration by JTC1/SC2/WG2 and UTC**

Date: **2017-08-25**

1. Background

WG2N4688 has only one reference, namely 藤花榭本, and is a proposal to encode the Small Seal Script by TCA and China. 藤花榭本 is the most qualified academically without any doubt. As known, there are lots of the series of Shuowen, but 大徐本 is the most important among them and is a reference model for compiling the summary dictionaries of the Chinese ancient scripts. Therefore, first encoding the Small Seal Script based on 藤花榭本 meets academic requirements. The other Small Seal characters found in other books could be encoded after encoding the characters on 藤花榭本 and thus, a new extended block for the other Small Seal characters without characters in 藤花榭本 can be developed.

2. Radicals Issue

There is an independent Kangxi Radicals block for CJK Han Script. For Small Seal Script, we have two alternatives to solve this issue.

(1) Create a new Small Seal Radicals block like Kangxi Radicals block.
(2) Use kRadicalNumber to tag the radical information as a new property for the Small Seal Script. This property doesn't apply to the characters which aren't as the radicals. Examples are as follows:

84	00834		𠘨	□	□	22	Zhengzhuan

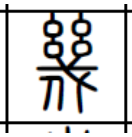
U+3xxxx kRadicalNumber 22 # ？

1	09207			糸	糸	467	Zhengzhuan

U+3xxxx kRadicalNumber 467 # ?

3. kRSunicode Issue

The Small Seal Script is taken as he classical forms of the CJK Han Script and is easily to consider have the radical information, the strokes information, even the first stroke information and the total strokes information. Among them, the radical information is very important and necessary, we think, but others are not. There is no rigorous disciplinary rule for the strokes counts of the Small Seal characters for the Chinese ancient scripts grammatology and the Chinese calligraphy. Therefore, we suggest use kRUnicode or kRShuowen to tag the radical information for each Small Seal character. It's better and more operable than using kRSUnicode to tag directly. Examples are as listed follows:

468	02766			幾	幺	124	Zhengzhuan

U+3xxxx kRUnicode 127 # ?

104	06804			馹	馬	370	Zhengzhuan

U+3xxxx kRUnicode 370 # ?

Some characters are followed by one or more variant forms. The radicals of these variant forms are as the same as their corresponding Zhengzhuan customarily. For example, the radical of "箕" is "竹", and "其" is the corresponding variant form of "箕", so the radical of "其" is "竹" as well. For the convenience of use, there will be a slight difference between the kRUnicode value for "箕" and "其".

165	03335			箕	箕	144	Zhengzhuan

U+3xxxx kRUnicode 144 # ?

169	03339						Variant

U+3xxxx kRUnicode 144' # ?

In addition, the radical information of “其” can’t be found out by its glyph directly, so we suggest add the second kRUnicode value for “其”. Other cases of variant forms are dealt with this way.

4. Collation Issue

The ranking method of Shuowen Jiezi is a common method to compile the summary dictionaries of other Chinese ancient scripts. We suggest use this method to rank all the Small Seal Characters in this block.

When some possible new Small Seal characters not included in 藤花榭本 will be found out and need to encode in the future, we can rank them by this ranking method. The characters which the radicals are the same could refer to the code points of their corresponding CJK Han characters to rank.

Number of Line

5. Suspected Repetitive Characters Issue

Suzuki-San pointed out the suspected repetitive characters on his WG2 documents. We have found out more cases like this, please see the following:

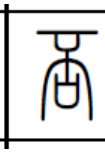
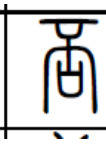
00939 vs 02078 (右), 00980 vs 03407 (吁), 02780 vs 04307 (敖), 00888 vs 06060 (吹), 03436 vs 07327 (愷), 08060 vs 09810 (邍), 01005 vs 08352 (否)

In these cases, there are a slight differences between the original glyphs of 00939 vs 02078 (右) and 08060 vs 09810 (邍), so we suggest to disunify them.

The glyph of 00980 vs 03407 (吁), 02780 vs 04307 (敖), 00888 vs 06060 (吹), 03436 vs 07327 (愷) and 01005 vs 08352 (否) are exactly the same. We suggest unify them and keep the code points for the character which is unified by the other one empty.

The encoded characters should be added the second kRUnicode values. Examples are as follows:

01005 will be unified with 08352.

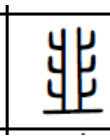
255	01005			否	口	22	Zhengzhuan
6	08352			否	不	432	Zhengzhuan

U+3xxxx kRUnicode 432;22 # ？

6. The Original Volume and Page Information Issue

We suggest create four properties which are kTHXVolumePage, kCCZVolumePage, kPJGVolumePage, kJGGVolumePage to record the original volume and page information on **藤花榭本**, **陳昌治本**, **平津館孫本** & **汲古閣本**. The format of the record properties is "vvv.pp" which the initial two "v"s means the volume number, the third "v" means "**卷上**" or "**卷下**" and "pp" means the page number.

kTHXVolumePage: [0-9]{3}\.[0-9]{2}
kCCZVolumePage: [0-9]{3}\.[0-9]{2}
kPJGVolumePage: [0-9]{3}\.[0-9]{2}
kJGGVolumePage: [0-9]{3}\.[0-9]{2}
Example is as follow:

408	05018			韭	韭	266	Zhengzhuan

藤花榭本：卷七下第一頁，陳昌治本：卷七下第二頁，平津館孫本：卷七下第一頁，汲古閣本：卷七下第二頁

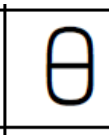
U+3xxxx kTHXVolumePage 072.01 # ?
U+3xxxx kCCZVolumePage 072.02 # ?
U+3xxxx kPJGVolumePage 072.01 # ?
U+3xxxx kJGGVolumePage 072.02 # ?

7. Character Name Issue

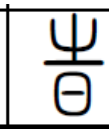
We suggest use SN on WG2N4688 to name the characters.

SMALL SEAL-[0-9]{5}|X\[0-9]{3}

Examples are as follows:

1	04638			日	日	231	Zhengzhuan

SMALL SEAL-04638

5	04642			眚	日	231	Variant

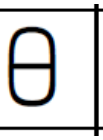
SMALL SEAL-04642

77	X142				日	231	Newly added character

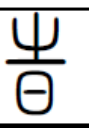
SMALL SEAL-X142

8. UCD Issue

Horizontal and vertical writings should be used in the future Small Seal texts. The horizontal writing needs to be left-to right and right-to-left, the vertical writing needs to be top-to-bottom. So we suggest the Bidirectional Class Values should be “ON” and they are different from this property of CJK Han script. Examples are as follows:

1	04638			日	日	231	Zhengzhuan

3xxxx;SMALL SEAL-04638;Lo;0;ON;;;;;N;;;;;

5	04642			昔	日	231	Variant

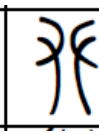
3xxxx;SMALL SEAL-04642;Lo;0;ON;;;;;N;;;;;

77	X142				日	231	Newly added character

3xxxx;SMALL SEAL-X142;Lo;0;ON;;;;;N;;;;;

9. kTranscription Issue

Each Small Seal characters should have their one or more corresponding transcribed CJK Han characters theoretically. We suggest create the kTranscription property to record the information. Example is as follow: 不是需要的屬性。

387	01910			兆	𠂤	63	Zhengzhuan

U+3xxxx kTranscription U+20B1C □;U+7676 𠂤 # ？

10. kZhengzhuan Issue

We suggest tag whether the character is Zhengzhuan by the kZhengzhuan property. Examples are as follows:

371	01894			彝	奴	62	Zhengzhuan

U+3xxxx kZhengzhuan Yes # ？

372	01895			算			Variant

U+3yyyy kZhengzhuan No:U+3xxxx # ？

11. Create the Small Seal Database or SmallSealSource.txt

We suggest create the Small Seal Database or SmallSealSource.txt to record and preserve all the properties we mentioned above, such as kRadicalMumber, kRUnicode, kTHXVolumePage, kCCZVolumePage, kPJGVolumePage, kJGGVolumePage, kTranscription and kZhengzhuan.

一、背景

WG2N4688 中的參考來源版本只有藤本。若從學術意義上說，藤本無疑是最合適的。眾所周知，說文系的書籍汗牛充棟，但大徐本是其中最為重要的，也是後世文字編編纂參考的範本。因此先以藤本作為底本進行編碼是符合学术要求的。在其他字書中存見的小篆字符可以在這次編碼完成後再作收集整理，形成新的小篆增補區。

二、說文部首的處理

CJK 漢字有獨立的康熙部首區塊，對於說文部首，可以有兩種選擇方案。

（1）像康熙部首區塊一樣，建立一個小篆部首區塊，區塊名稱如 Small Seal Radicals。

（2）使用 kRadicalNumber 進行屬性標記。(部首字以外的字不包含這個屬性。) 舉例如下：

84	00834			口	口	22	Zhengzhuan

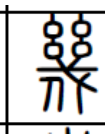
U+3xxxx kRadicalNumber 22 # ?

1	09207			糸	糸	467	Zhengzhuan

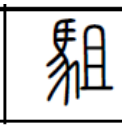
U+3xxxx kRadicalNumber 467# ?

三、kRSUnicode 問題

小篆作為漢字的古典形式 (classical form)，容易被認為和 CJK 漢字一樣需要有歸部和部外筆劃，甚至部外首筆、整字總筆畫等資訊。歸部對於小篆而言非常重要，而且必須，但部外筆劃等並非必要。文字學意義上並沒有對小篆的筆劃數有嚴格的學科規則，書法意義上有一定的規則，但那是後起規則，只為了毛筆書寫小篆的方便，並非小篆本身應有的屬性。因此，僅僅記錄 kRUnicode 或 kRShuowen 比記錄 kRSUnicode 更可行。舉例如下：

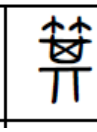
468	02766			幾	幺	124	Zhengzhuan

U+3xxxx　kRUnicode 127#　?

104	06804			馹	馬	370	Zhengzhuan

U+3xxxx　kRUnicode 370#　?

在說文的體系中，部分字有重文。按照學術習慣，重文也一樣歸於正篆的部首下。如作為正篆的“箕”歸部是竹部，作為重文的“其”歸部也應是竹部，但為了使用方便，可以稍作區別。

165	03335			箕	箕	144	Zhengzhuan

U+3xxxx　kRUnicode 144#　?

169	03339						Variant

U+3xxxx　kRUnicode 144'　#　?

另外，“其”字在字形上不能直接看出竹的歸部信息來，因此可以考慮增加第二歸部，這樣可以方便於使用者查找。其他重文、古文等仿此。

四、排序問題

學界編輯古文字辭書常用的是說文部首排序，在編碼層面上對說文部首排序的全面繼承也符合學術習慣。若需要在完成這次編碼後增補新的小篆字頭，或以後用於其他古文字的編碼，可以考慮先按说文部首排序，同一部首內的字按照相應隸定字的 Unicode 先後排序。

五、疑似重複字問題

鈴木先生在其 WG2 的文件中指出有重複字問題。我們發現的疑似重複字，包括如下部分：

00939 vs 02078 (右),

00980 vs 03407 (吁),

02780 vs 04307 (敖),

00888 vs 06060 (吹),

03436 vs 07327 (愷),

08060 vs 09810 (澨),

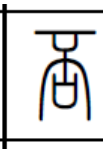
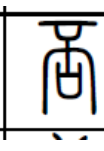
01005 vs 08352 (否)

其中 00939 和 02078（右）、08060 和 09810（澨）這兩對的字形是稍有區別的，我們建議不對這兩對進行認同。

00980 和 03407（吁）、02780 和 04307（敖）、00888 和 06060（吹）、03436 和 07327(愷)、01005 和 08352(否)這五對在字形和字義上都看不出有區別，我們建議只保留其中一個，並保留被認同字符的原有位置，暫不做任何安排。

保留下來的字符，可以考慮增加第二歸部。舉例如下：

01005 與 08352 認同。

255	01005			否	口	22	Zhengzhuan
6	08352			否	不	432	Zhengzhuan

U+3xxxx　kRUnicode 432;22 #　?

六、原書卷頁信息

建立 kTHXVolumePage、kCCZVolumePage、kPJGVolumePage、kJGGVolumePage 等分別標記藤花榭本、陳昌治本、平津館孫本、汲古閣本中的原書卷頁信息，以 vvv.pp 的格式記錄。前两个 v 指的是卷数，第三个 v 指的是该卷的“卷上”或“卷下”，pp 指的是页码。

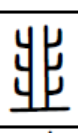
kTHXVolumePage: [0-9]{3}\.[0-9]{2}

kCCZVolumePage: [0-9]{3}\.[0-9]{2}

kPJGVolumePage: [0-9]{3}\.[0-9]{2}

kJGGVolumePage: [0-9]{3}\.[0-9]{2}

舉例如下：

408	05018			韭	韭	266	Zhengzhuan

藤花榭本：卷七下第一頁，陳昌治本：卷七下第二頁，平津館孫本：卷七下第一頁，汲古閣本：卷七下第二頁

U+3xxxx kTHXVolumePage 072.01 # ?

U+3xxxx kCCZVolumePage 072.02 # ?

U+3xxxx kPJGVolumePage 072.01 # ?

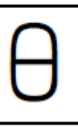
U+3xxxx kJGGVolumePage 072.02 # ?

七、字符名稱

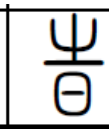
按照 WG2N4688 的序列號來標記。

SMALL SEAL-[0-9]{5}|X\[0-9]{3}

舉例如下：

1	04638			日	日	231	Zhengzhuan

SMALL SEAL-04638

5	04642			眚	日	231	Variant

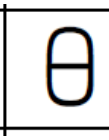
SMALL SEAL-04642

77	X142				日	231	Newly added character

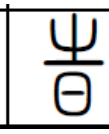
SMALL SEAL-X142

八、UCD 問題

小篆的使用應有橫排、豎排可用，橫排也應有從左到右和從右到左兩種可能，因此 UCD 屬性中 Bidi Class 一欄可以考慮採用 ON，而不與 CJK 漢字所使用的 L 相同。舉例如下：

1	04638			日	日	231	Zhengzhuan

3xxxx;SMALL SEAL-04638;Lo;0;ON;;;;;N;;;;;

5	04642			眚	日	231	Variant

3xxxx;SMALL SEAL-04642;Lo;0;ON;;;;;N;;;;;

77	X142				日	231	Newly added character

3xxxx;SMALL SEAL-X142;Lo;0;ON;;;;;N;;;;;

九、kTranscription 問題

每個小篆字頭有一個甚或多個可以對應的 CJK 字，因此需要有專門的屬性進行

記錄。舉例如下：

387	01910			𠂤	𠂤	63	Zhengzhuan

U+3xxxx kTranscription U+20B1C 𠂤;U+7676 𠂤 # ?

十、kZhengzhuan 問題

對正篆與重文信息進行標記。舉例如下：

371	01894			彧	收	62	Zhengzhuan

U+3xxxx kZhengzhuan Yes # ?

372	01895			彧			Variant

U+3yyyy kZhengzhuan No:U+3xxxx # ?

十一、建立 Small Seal Database 或 SmallSealSource.txt

以上所談及的 kRadicalMumber、kRUnicode、kTHXVolumePage、kCCZVolumePage、kPJGVolumePage、kJGGVolumePage、kTranscription、kZhengzhuan 等需要用 Small Seal Data 或 SmallSealSource.txt 這樣的載體來記錄與保存。

(End of Document)