Universal Multiple-Octet Coded Character Set UCS

IRG N1842
Date: 2012-02-10

Source:	Japan
Title:	Japanese proposal to the agenda for Old Hanzi Tokyo meeting
Meeting:	Old Hanzi ad-Hoc Tokyo, 2012-02-20/23
Status :	
Actions required	For discussion at Old Hanzi ad-Hoc Tokyo, 2012-02-20/23 Electronic
Distribution:	IRG, Old Hanzi group
Medium :	
Page:	

In this document Japan proposes the following prioritization of the agenda items assigned to Old Hanzi group in IRG#37. The member bodies are encouraged to submit the feedback on the following prioritized agenda items. By the submitted feedback, Old Hanzi group will be able to finish the agenda items assigned by IRG within the meeting.

A) The clarification of the bibliography list in Old Hanzi P&R version 3 draft, section 3.7.
B) The clarification of the indexing method of the glyphs which do not have corresponding SW and UCS characters.
C) The consideration of the coverage of encoded UCS characters that can be used in the column "Corresponding Modern Character (UCS)".
D) The completion of the definition of all the entries required of the submission.
E) IRG M37.9: the design of Old Hanzi coding framework.

The most time-consuming item would be E), IRG M37.9; the description how Oracle Bone scripts to be coded. The specifications of the requirement to be supported are also needed; which features the coded Oracle Bone must guarantee.
Japan thinks the interoperability with existing resources is most important. The relationship with previous Japanese proposal IRG N1771 is described in the appendix of this document.

Appendix

A.1. Action Items assigned by IRG

The action items assigned by IRG are following. For detail, please check IRG N1810 (IRG#37 meeting resolution), IRG M37.8, and IRG N1827 (the report of Old Hanzi discussion in IRG#37).

A) The clarification of the bibliography list in Old Hanzi P&R section 3.7.

B) The clarification of the indexing method of the glyphs which do not have corresponding SW and UCS characters.

C) The consideration of the coverage of encoded UCS characters that can be used in the column "Corresponding Modern Character (UCS).
   - In IRG#36 at Chongqing, Old Hanzi group agreed to restrict the corresponding UCS characters to URO.
   - IRG experts had pointed out a possibility that SW character should be mapped to Ext B, C, D characters.

D) The completion of the definition of all the entries required of the submission.

E) IRG M37.9: the design of Old Hanzi coding framework.
   Old Hanzi group is requested to describe how Oracle Bone scripts to be coded, with the specifications of the requirement to be supported. After summarizing the requirement how Oracle Bone scripts should be coded, Oracle Bone group should review IRG N 1771 (Japanese proposal of coding framework) and make the feedback to it.

A.2. Comment from Japan to the assigned work items

In following, comments from Japan to each work items are described.

A) The clarification of the bibliography list in Old Hanzi P&R section 3.7

Japan had already submitted the draft including the bibliography information. Because all references are submitted by China and TCA, so Japan cannot complete it. Japan request the review of the reference lists and give the completion of the list.

B) The clarification of the indexing method of the glyphs which do not have corresponding SW and UCS characters.

Japan still think the natural classification in Yinxu Jiag Keci Leizuan (殷墟甲骨刻辭類纂, LZ) is the best sorting method for Oracle Bone, thus such characters (no Shuowen or no UCS characters can be corresponded) should be ordered by it. A recent derivative of LZ, Xinbian Jiaguwen Zixing Zongbiao (新編甲骨文字形總表, ZB) can be an option to sort the glyphs missing in LZ.

If such natural classification is unacceptable for non-Shuowen characters and the reason to follow Shuowen is the popularity (the number of Oracle Bone dictionary using Shuowen order is larger than others), Japan propose to investigate the existing Oracle Bone dictionaries using Shuowen order, and

choose 1 dictionary to follow.

To choose a dictionary to follow, the most important points to be cared are:

- The dictionary should not show single glyph at multiple places. It is to minimize the ambiguity of the location of glyph.
- The source materials of the dictionary are similar with that in Old Hanzi database. It is to minimize the difference of the glyph collection.
- The source information is convertible with that in Old Hanzi database. It is to minimize the error of the source information translation.

Considering these issues, clearly 甲骨文編 (WB) should not be followed although it can be one of the most referred dictionary; its source information was before the publishing of 甲骨文合集 (HJ), and it is difficult to convert with its source information to the Old Hanzi database. 甲骨文字典 (ZD) had taken some glyphs from JH, but most glyphs were taken from the references published before JH.

新甲骨文編 (SJWB) would be the most considerable candidate, because most parts are taken from JH.

If the same ordering should be applied to Jinwen and Xiaozhuan, the dictionary covering such materials should be considered; like 古文字類編, 漢語古文字字形表, etc. But their source information is pre-HJ style and there is conversion difficulty. Also the coverages of the Oracle Bone glyph in these dictionaries are slightly questionable. At present, Japan could not find appropriate candidates whose source information are convertible with current Old Hanzi database.

C) The consideration of the coverage of encoded UCS characters that can be used in the column "Corresponding Modern Character (UCS).

To consider the inclusion of CJK Unified Ideograph Extension A—D, the digitized version of Shuowen must be fixed for first. Especially CJK Unified Ideograph Extension B includes many "Guwen" glyphs taken from Kangxi. They can be used to transliterate Xiaozhuan glyphs, and uncontrolled usage of them makes it difficult to sort the glyphs to Shuowen order by UCS characters.

Thus, before the discussion to extend the coverage of acceptable UCS characters to relate with Oracle Bone, the stabilization of digitized Shuowen text is needed. Referring "Shuowen Daxu" is insufficient. For example, an edition of Shuowen Daxu (說文眞本, based on 汲古閣本) used U+26903 "皇" instead of "皇". If "皇" (U+7687) is preferred, modern typesetted version must be specified, at least. The best solution would be reference to the existing digitized text, to prevent the error by manual retyping of the text.

D) The completion of the definition of all the entries required of the submission.

The distinction of "image", "shape", "glyph" and "character" was most important issue commented in IRG#37. Japan thinks the data in "original shape/glyph" is apparently an "image", because it includes some noise. The column "imitation shape/glyph" is defined as "truly trace" since IRG#37,

thus it would be an "image" or "shape", because the difference from the "original shape/glyph" is only removal of noise, connecting the cracked lines, etc.

E) IRG M37.9: the design of Old Hanzi coding framework.

Japan already proposed a coding framework in IRG N1771. The background idea of the coding framework proposed by Japan is following.

The interoperability with existing Oracle Bone data (published concordances, glyph/character charts, dictionaries, digital databases) is most important for an industrial standard. Especially, ISO/IEC 10646 is maintained to be a standard with no corrigendum cancelling previously coded character.
There are several Oracle Bone fonts, but most of them seem to be based on the images scanned from WB, LZ (e.g. CDP project in Sinica) or ZB (e.g. CHANT project in Chinese University in Hong Kong). Considering that Cuneiform standardization had selected the cross section of 2 different dictionaries to prevent an insertion of uncancellable error to ISO/IEC 10646, Japan decided to propose the stable cross section of existing resources. Unfortunately, the source information syntax are varied in each collection and difficult to compute the cross section without human error, Japan proposed the core set by counting the numbers of objects given in LZ (WB and ZB did not define the frequency of the characters). If the number of materials is more than dozens, it would be expected to be included in other collections (concordances, glyph charts, dictionaries, databases). But if the number of materials is less than 10, there is a possibility that LZ heading glyph is not included, or is unified with other glyphs in other collections, thus we excluded from the proposal.

(end of document)