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

Doc Type:	**Ideographic Rapporteur Group Document**
Title:	**Request to clarify some IDS issues**
Source:	**Eiso Chan (陈永聪, Culture and Art Publishing House)**
Status:	**Individual Contribution to IRG #53, Shenzhen, PRC**
Action:	**For consideration by IRG**
Date:	**2019-10-24**

In the IRG WS works and the Jianzi encoding works, there are some unclear issues related to IDS, so I request IRG to clarify them.

1. Scope of Character Description Component (CDC)

In IRG PnP, Glossary, it shows "*It includes all coded CJK unified ideographs, Kangxi Radicals, CJK Radical Supplements, and coded CJK Compatibility ideographs.*" Coded CJK unified ideographs mean all encoded characters in URO and all CJK extensions, Kangxi Radicals mean all encoded characters in *Kangxi Radicals* block (U+2F00 to U+2FDF), CJK Radical Supplements mean all encoded characters in *CJK Radical Supplement* block (U+2E80 to U+2EFF) and coded CJK Compatibility ideographs mean all encoded characters in *CJK Compatibility Ideographs* block and *CJK Compatibility Ideographs Supplement* block.

When we check the script property values in UCD for the characters mentioned above as below, we will find that they are related to Han.

Characters	**Script property values**
CJK Unified Ideographs	Han
Kangxi Radicals	Han
CJK Radical Supplements	Han
CJK Compatibility Ideographs	Han

However, there are also other Han characters in CJK Symbols and Punctuation block, they are the Ideographic Iteration Marks, Ideographic Number and the Suzhou Numerals shown as below.

UCS	**Character**	**Character Name**
U+3005	々	IDEOGRAPHIC ITERATION MARK
U+3007	〇	IDEOGRAPHIC NUMBER ZERO
U+3021	〡	HANGZHOU NUMERAL ONE
U+3022	〢	HANGZHOU NUMERAL TWO
U+3023	〣	HANGZHOU NUMERAL THREE

UCS	Character	Character Name
U+3024	〤	HANGZHOU NUMERAL FOUR
U+3025	〥	HANGZHOU NUMERAL FIVE
U+3026	〦	HANGZHOU NUMERAL SIX
U+3027	〧	HANGZHOU NUMERAL SEVEN
U+3028	〨	HANGZHOU NUMERAL EIGHT
U+3029	〩	HANGZHOU NUMERAL NINE
U+3038	〸	HANGZHOU NUMERAL TEN
U+3039	〹	HANGZHOU NUMERAL TWENTY
U+303A	〺	HANGZHOU NUMERAL THIRTY
U+303B	〻	VERTICAL IDEOGRAPHIC ITERATION MARK

In IRG WS2017 work, Việt Nam proposed the IDS for V-F1D09 (WS2017-00012) is ⿰⿰一⿰丨丿可, ROK also registered this character as KC-13265, but they had not submitted this character to IRG. In the v3.0 review cycle, I pointed out the Suzhou numerals are not allowed in the current IDS syntax, and IRG confirmed this in IRG #52. Lee provided his comment as "*Vietnam believes that all common element such as* 丨 *should be available as IDS to reduce long, confusing sequences. If the Suzhou numerals are not allowed, IRG should move to encode more components for IDS.*" (cf. https://hc.jsecs.org/irg/ws2017/app/index.php?id=00012)

For these two ideographic iteration marks, the newest code charts have shown U+3005 shares the same glyph shape with U+206A4 (𠂤) which is cited from the 748 code chart and the G-Source reference value has been changed to GFZ-FCA2 in WG2 N5114 according to IRGN2245, and, U+303B is not needed as any element for CJKUI as we known. For U+3007 (〇), it has not included in GB/T 2312—1980 and the first edition of CNS11643, but China added it to the standards since GBK as 0xA996 [3.2.8 and GBK/5 in GBK or IRGN278 "3.2.8. 汉字 "〇" (GB 13000.1/UCS=3007 "零")", 5.2.i in GB 18030—2000 and 5.2 in GB 18030—2005], CNS11643 has included it as 1-4243 since 2007 version, HK SAR has added it as 0xC6E2 since HKSCS-2001, Japan has added it as 01-27 since JIS C 6226-1978 or JIS X 0208-1978 (aka 0x815A in Shift_JIS), but not found in ROK, DPRK and Việt Nam standards. U+3007 could be treated as the special CJKUI. Notice that U+3007 (〇) is needed for one Jianzi fingering 𠂤 (從圈), which will be mentioned in Section 2 in this document.

I checked U+3022 and other Suzhou Numerals are needed for other different CJKUI and maybe they will be useful to encode Nanyin Notation characters in future.

On the other hand, the CJK strokes in CJK Strokes block, the fullwidth question mark (U+FF1F/？) and the CJK supplementary components mentioned in IRGN2225, IRGN2230 have been used widely in IDSes for IRG WS works. For the CJK supplementary components, it begins with &/U+0026, and ends with ;/U+003B, like the tradition of SGML/html/xml.

So, I request to update the scope of CDC as below in IRG PnP.

It includes all coded CJK **U**nified **I**deographs, Kangxi Radicals, CJK Radical Supplements, ~~and~~ coded CJK Compatibility **I**deographs**, CJK Strokes, Fullwidth Question Mark, Suzhou Numerals and CJK Supplementary Components**.

The syntax should be changed as below.

• IDS::=<Binary_Symbol><IDS1><IDS1>|<Ternary_Symbol><IDS1><IDS1><IDS1> • <IDS1> ::= <IDS> | <CDC> • <CDC> ::= coded_ideograph | coded_radical | coded_component **| coded_stroke | U+FF1F | supplementary_component | suzhou_numeral**

- <Binary_Symbol> ::= ⿰ | ⿱ | ⿲ | ⿳ | ⿴ | ⿵ | ⿶ | ⿷ | ⿸ | ⿹ | ⿺
- <Ternary_Symbol> ::= ⿲ | ⿳

If IRG has confirmed the CJK strokes, fullwidth question mark, the CJK supplementary components or the Suzhou numerals are not allowed for IDS in the IRG encoding works, it's better to clarify in IRG PnP, which will be helpful for the submitters and reviewers.

The following IDS syntax is cited from the latest version of Unicode Core Specification.

> ***Ideographic Description Sequences.*** Ideographic Description Sequences are defined by the following grammar. The list of characters associated with the Ideographic and Radical properties can be found in the Unicode Character Database. In particular, the Ideographic property is intended to apply to other siniform ideographic systems, in addition to CJK ideographs. *[...]*
> *IDS* := *Ideographic* | *Radical* | *CJK_Stroke* | *Private Use* | U+FF1F
> | *IDS_BinaryOperator IDS IDS*
> | *IDS_TrinaryOperator IDS IDS IDS*
> *CJK_Stroke* := U+31C0 | U+31C1 | ... | U+31E3
> *IDS_BinaryOperator* := U+2FF0 | U+2FF1 | U+2FF4 | U+2FF5 | U+2FF6 | U+2FF7 | U+2FF8 | U+2FF9 | U+2FFA | U+2FFB
> *IDS_TrinaryOperator*:= U+2FF2 | U+2FF3

In the Unicode IDS syntax, *Ideographic* means all the characters which their scripts are belong to *Han*.

2. Encoding Model for Jianzi

I, the source showed Culture and Art Publishing House, submitted WG2 N5041 to request to encode Jianzi in Unicode and ISO/IEC 10646 in future, and now this work is analyzing and ongoing. Prof. Tang, the source showed China National Database (aka Chinese Characters Repertoire), pointed out there are something not better in the encoding model based on musical notation meanings. I read Prof. Tang's document and discussed with the Guqin experts and designers in my team, we thank the comment, but we still don't think the encoding model based on IDS provided by Prof. Tang is well.

2.1. Unstable Sequences

After WG2 #68, Mr. Chen Zhuang, Prof. Tang Yingmin and I met at PKU to discuss the Jianzi encoding model. I pointed out why the IDS method is not better.

In Unicode core specification, it shows the important description as below. (p. 725 in Unicode, 12.0.0)

> The Unicode Standard does not define equivalence for two Ideographic Description Sequences that are not identical. *Figure 18-9* contains numerous examples illustrating how different Ideographic Description Sequences might be used to describe the same ideograph. In particular, Ideographic Description Sequences should not be used to provide alternative graphic representations of encoded ideographs in data interchange. Searching, collation, and other content-based text operations would then fail.

In IRG PnP, Annex B, it shows both of ⿰亻頃 and ⿰化頁 should be acceptable for the IDS for 傾 (U+50BE) and so on.

For the big-sized characters in the Jianzi system, there are also the similar situations. For example, the IDS for 芶 (散勾三弦) could be used ⿱艹⿰勹三 or ⿱⿱艹勹三 as their IDS. The first one is easy for the end user if he or she doesn't know Guqin, but more Guqin players will

choose the second one. Prof. Tang asked us if it's better to treat 散 as the left hand fingering, I discussed this issue with our chief adviser Prof. Wu Zhao and other Guqin experts, and their comments are 散 is not suitable as the left hand fingering.

The other sample is 𱻬 (食指四徽绰注勾剔四弦). The elements for this Jianzi big-sized character are 亻 (possibly unified by U+4EBB/亻), 四 (possibly unified by U+56DB/四), ⺊ (possibly unified by U+2E8A/⺊ or add a new character), 氵 (possibly unified by U+6C35/氵), 冓 (no similar or unifiable character), 四(possibly unified by U+56DB/四). If we need to use the IDS as the encoding model, it will make the sequence hard to understand (like ⿲⿰亻四⿲⿰⺊冓氵四). Notice that 冓 is a single fingering for Guqin, so it must be treated as a single element in any form of encoding model. And sometimes, we will also need ⿲, one of the unencoded IDC mentioned in IRGN2273R=L2/18-012 but not accepted by UTC.

2.2. Information confusion

During the discussion, some WG2 experts consider to unifying Jianzi fingering which share the same or similar glyph shapes with CJKUI to the corresponding CJKUIs, eg. unify 勹 with 勹 (U+52F9), unify 十 with 十 (U+2C09B), and then, encode the disunifiable ones in other way. We think this method is worth for further discussion. If we allow the IDC in the Jianzi encoding model, the machine is hard to distinguish Jianzi from CJKUI, but it's necessary to make them have something different in the information interchange.

2.3. Vendors' burdens

Andrew West and I once used IDS and `ccmp` GSUB feature to make the Jianzi encoding model in a font.

In Windows, the visual glyphs can display in Microsoft Word when we change the ligature option for OpenType, but it can't be printed to the PDF file; they can't display in other apps.

In Mac, the font can run regularly.

I think the original IDS form which show the characters and the IDCs together is useful as UTC wrote as above, and I will provide other encoding model for Jianzi big-sized characters / clusters in the coming proposal to WG2 and UTC with other experts.

All in all, I request IRG to clarify the IDS target is the current performance and it's not best to make them to do anything else.

3. U+303E issue

We have used IDEOGRAPHIC VARIATION INDICATOR (〾, U+303E) in the IDSes in the IRG WS since IRG WS2015, but there is no any explanation in IRG PnP.

In Unicode specification, it shows the following.

> ***Interaction with the Ideographic Variation Mark.*** As with ideographs proper, the Ideographic Variation Mark (U+303E) may be placed before an Ideographic Description Sequence to indicate that the description is merely an approximation of the original ideograph desired. A sequence of characters that includes an Ideographic Variation Mark is not an Ideographic Description Sequence.

On the other hand, UTC will update the definition as below in Unicode, 13.0.

> ***Interaction with the Ideographic Variation Indicator.*** U+303E ideographic variation indicator (IVI) normally occurs before a CJK unified ideograph, but it may also be placed before an Ideographic Description Sequence to indicate that the description is merely an

approximation of the ideograph desired. The IVI is not considered a part of the Ideographic Description Sequence and does not invalidate the sequence.

I request for consideration to add this paragraph in IRG PnP, Glossary as well.

4. Additional Issue

Yifan pointed out the so-called “para-ideographs” in L2/19-346 to the Script Ad-hoc Group and UTC, but there is no comment in the Script Ad-hoc Group meeting in Sept. 27th, 2019, see L2/19-343.

We reviewed this document which raises a number of points about the Gongche characters and other “para-ideographic” characters. Because the Script Ad Hoc reviews proposals for non-CJK characters, we recommend the document be submitted to the UTC and to IRG for discussion.

As the original submitter of the Gongche proposal, I oppose to move the seven Gongche characters out of the end of CJK Ext. B to destabilize Unicode, Version 13.0, but it’s necessary for all the IRG expert to consider how to encode other so-called “para-ideographs” in Yifan’s document.

I understand the “para-ideographs” means the hybrid characters and the characters used in the ideograph running text but the stroke shapes are not like the common ones. In the list provided by Yifan, the first character in Table 2 could be unified with U+211A0 (𑋠), which has been added to BabelStoneHan via `ss01` and `cv01`; the first character in Table 5 could be unified with U+2CF36 (𬼶) because of cognate; the second character in Table 5 is the cursive form of U+2E574 (𮕴).

Andrew West and I also collected more hybrid character like Yifan said. Andrew added them to his famous font BabelStoneHan in PUA.

PUA	Char.	Pseudo-IDS	Note
U+F3E2	庴	⿰广⿰タカ	Kanji-Katakana hybrid = 鷹 taka in the Tokyo place name 墨砧～ = ボク きぬた たか (Boku kinuta taka).
U+F8C0	杚	⿰木 K	Sawndip-Latin hybrid = ge (?) "pine"
U+F8C1	纪	⿰纟 K	Sawndip-Latin hybrid = gej "to untie" (= 擘)
U+F8C2	耆	⿰老 K	Sawndip-Latin hybrid = geq "old" (= 耆)
U+F8C3	疴	⿰疒 A	Sawndip-Latin hybrid = ae "cough" (= 痎)
U+F8C4	丫	⿰丫 E	Sawndip-Latin hybrid = ngez (?) "branch"
U+F8C5	躬	⿰身 N	Sawndip-Latin hybrid = enj "to stick out one's chest or stomach" (= ⿰益先 GZ-1482101 in Ext. G)
U+F8C6	苊	⿰艹 M	Sawndip-Latin hybrid = em "kind of grass" [芭芒] (= 菁)
U+F8C7	囤	⿰口도	Hanja-Hangul hybrid = 圖 do "map"

PUA	Char.	Pseudo-IDS	Note
U+F3E2	庴	⿰广⿰タカ	Kanji-Katakana hybrid = 鷹 taka in the Tokyo place name 墨砧～ = ボク きぬた たか (Boku kinuta taka).
U+F8C8	庇	⿰广ふ	Kanji-Hiragana hybrid = 腐 in the word 豆腐 (とふ tofu)
U+F8C9	庘	⿰广K	Kanji-Latin hybrid = 慶 in the name of Keiō University (慶應大学)
U+F8CA	庁	⿰广O	Kanji-Latin hybrid = 應/応 in the name of Keiō University (慶應大学)
U+F8CB	庙	⿰广⿰KO	Kanji-Latin hybrid = 慶應/慶応 in the name of Keiō University (慶應大学)

In the current CJK encoding method, the hybrid characters mentioned above can't be encoded in CJKUI, but it's not better to encode in another new block. It's hard to use IDS to describe based on the current IDS syntax.

I request the IRG experts to consider how to handle these hybrid characters.

5. Acknowledgement

Thanks for the feedback comments from Andrew West (魏安), Lee Collins (康立論), Dr. Ken Lunde (小林劍), William Nelson, John Knightley (李忠仕), Kushim Jiang (姜兆勤) and Henry Chan (陳輝恒).

The new definition of IVI is written by Ken Whistler.

(End of Document)