

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: Feedback on IRGN2492 and the preliminary encoding method of early Chinese organic chemical character, Sanskrit transcription, Tibetan transcription, Tangut transcription and Jianzi Musical Notation
Source: Eiso Chan (陈永聪, Culture and Art Publishing House)
Status: Individual Contribution to IRG #57, online meeting
Action: For consideration by IRG
Date: 2021-09-07

Dr. Ken Lunde and John Jenkins submitted [L2/21-118](#) Preliminary proposal to add a new provisional *kIDS* property (*Unihan*) to UTC, which was reviewed by [CJK & Unihan Group](#). Yi Bai provided his comments on the component list in [L2/21-134](#) Collections of ideograph components for use in *IDSes* which was also reviewed by CJK & Unihan Group. And then, Ken and John revised their document as [L2/21-118R](#) and submitted to IRG as [IRGN2492](#). I provided my feedback on *kIDS* property and proposed the encoding method for the ideographic complex script(s) preliminarily in this document.

1. New IDCs

In L2/21-118(R), the authors suggested 5 new IDCs. 4 of them had been suggested by Tao Yang, Yifan Wang and Me in [IRGN2273R](#). And Kushim Jiang provided his [feedback](#) on the unary IDCs.

1.1. IDC-SFR and -SLR

In fact, the IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM RIGHT (SFR) and the IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER RIGHT (SLR) are very necessary to use for *IDSes*, and we have provided some useful examples in [IRGN2273R](#). Please see Table 1.1.

Table 1.1. IDC-SFR and -SLR

IDC-SFR	IDC-SLR
┌ - - ┐ ┌ - ┐ └ - ┘ └ - ┘	┌ - ┐ ┌ - ┐ └ - ┘ └ - ┘

1.2. IDC-HRE and -ODR

The pictures of the IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL REFLECTION (HRE) and the IDEOGRAPHIC DESCRIPTION CHARACTER ONE HUNDRED EIGHTY DEGREE ROTATION (ODR) are shown in Table 1.2. The main comment on Kushim’s feedback is that he concerned if two or multiple of two consecutive same unary IDCs before one CJKUI or IDS would make the IDS was equal to the CJKUI or IDS itself. Please see Table 1.3. But the IDS with two of multiple of two consecutive different unary IDCs would not be equal to the IDS without these IDCs. Please see Table 1.4.

Table 1.2. IDC-HRE and -ODR

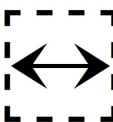
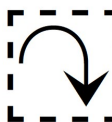
IDC-HRE	IDC-ODR
	

Table 1.3 IDS with two consecutive same unary IDCs

HRE, HRE, IDS = IDS
ODR, ODR, IDS = IDS

Table 1.4 IDS with two consecutive different unary IDCs


HRE, ODR, IDS ≠ IDS
ODR, HRE, IDS ≠ IDS

Kushim suggested moving HRE and ODR to IDMs (Ideographic Description Mark, and this term has not been used in the current standard yet). After the discussion in CJK & Unihan Group meeting for UTC #168, I think there is no need to use the term IDM, but we need to clearly recognize the situations mentioned in Table 1.3 are important for the future encoding review works, that means we can’t prohibit the IRG submitters, even the end users, use the unary IDCs as Table 1.3 showed, but the IDS checking program should make the IDS with two consecutive same unary IDCs and the IDS without two consecutive same unary IDCs equivalent.

1.3. IDC-SS

The picture of the IDEOGRAPHIC DESCRIPTION CHARACTER STROKE SUBTRACTION (SST) is shown as below. Please see Table 1.5.

Table 1.5 IDC-SST

IDC-SST


The examples with this IDC mentioned in IRGN2492 could also use other IDCs, but this IDC

could make the IDS shorter, which will be better for the IRG encoding works in future. Please see Table 1.6.

Table 1.6 Compare for the IDS with IDC-SS and without IDC-SS

UCS & Char.	IDS with IDC-SST	IDS without IDC-SST
U+2002A 其	一其ノ (3)	𐤀𐤁𐤂𐤃𐤄𐤅 (6)
U+2002B 其	一其、 (3)	𐤀𐤁𐤂𐤃𐤄𐤅 (6)
U+2CEBB 豸	一豸ノ、 (5 or shorter)	𐤀𐤁𐤂𐤃𐤄𐤅 (5 or longer)






For the above three IDSeS with IDC-SST will be easier to understand by the reviewer than the corresponding ones without IDC-SST.

1.4. Abbreviations of IDCs

We will have 17 IDCs when these five are accepted in UCS and Unicode Standard, but it's hard to call them orally now. Therefore, I suggest using the abbreviations of IDCs as below. I suggest using three ASCII / Basic Latin letters to indicate the IDCs, and we should use the "IDC-XXX" form when we need to talk about them outside the IDS environment. Note that the abbreviation of U+303E (𐤀) is IVI.

Table. 1.7 Abbreviations of IDCs

UCS	Char.	Char. Name	Suggested Abbr.
U+2FF0	𐤀	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT	(IDC-)LTR
U+2FF1	𐤁	IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW	(IDC-)ATB
U+2FF2	𐤂	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT	(IDC-)LMR
U+2FF3	𐤃	IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW	(IDC-)AMB
U+2FF4	𐤄	IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND	(IDC-)FSR
U+2FF5	𐤅	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE	(IDC-)SFA
U+2FF6	𐤆	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW	(IDC-)SFB
U+2FF7	𐤇	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT	(IDC-)SFL
U+2FF8	𐤈	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT	(IDC-)SUL
U+2FF9	𐤉	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT	(IDC-)SUR
U+2FFA	𐤊	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT	(IDC-)SLL

UCS	Char.	Char. Name	Suggested Abbr.
		IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM RIGHT	(IDC-)SFR
		IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER RIGHT	(IDC-)SLR
		IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL REFLECTION	(IDC-)HRE
		IDEOGRAPHIC DESCRIPTION CHARACTER ONE HUNDRED EIGHTY DEGREE ROTATION	(IDC-)ODR
		IDEOGRAPHIC DESCRIPTION CHARACTER STROKE SUBTRACTION	(IDC-)SST

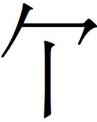
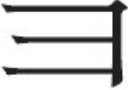

2. Component List



In L2/21-118, the authors provided a list of 60 components based on the 2021-05-26 version of IDS.TXT at BabelStone. And Yi Bai merged different component lists and suggested unifications. I provide my comments on the components as below.

2.1. Unnecessary components

The following components are unnecessary to encode in the standard. Please see Table 2.1. In this table, BS Syntax means the numbers mentioned in L2/21-118 and IDS.TXT at BabelStone, YB No means the numbers mentioned in L2/21-134.

Table 2.1 Unnecessary encoded components




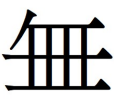


BS Syntax	YB No	Picture	Note
{09}	50		All the characters mentioned in examples column in L2/21-118 are the Vietnamese Nôm characters. For the rationale, all of them could also be “equivalent” to the CJKUIs which are the same as the left or the lower left element as the base character with U+16FF0 (𠄠). The code chart has showed U+16FF0 (𠄠) is derived from U+4E2A (个) and U+4E87 (𠄠), and the glyph shape is more similar to U+4E87 (𠄠). It’s better to use U+4E87 (𠄠) in IDS.
{41}	105		It’s better to use UTC-01005 or GZ-4591101 form which will be disunified from U+5F50 (𠄠) requested in IRGN2509 (L2/21-152).
{42}	106		It’s better to use U+5F50 (𠄠) when IRGN2509 is accepted.

BS Syntax	YB No	Picture	Note
{43}	107		It's better to use U+2E95 (𠄎).
{63}	65		The example characters mentioned under this component should be modified, because this component is the wrong form of U+821F (舟). I hope Vietnam NB should provide a document to confirm this one later.

2.2. Some components are also used as the fingering letters of the Jianzi Musical Notation

This part is out of the scope of IRG now, but during my research and the feedback comments from other WG2 and UTC experts, I need to care about the mappings between CJKUI and the basic Jianzi fingering letters, so I list the possible useful information as below.

Table 2.2 Components which are shared as Jianzi Fingering Letters

BS Syntax	YB No	Picture	Jianzi usage
{02}	66		喚 and so on
{03}	67		換, 喚, 換, 免, 換音 and so on
{47}	74		搗
{56}	76		無
	4		再
	20		緊

2.3. Code points of the new ideograph components

In the end paragraph of the part “New Ideograph Components” in IRGN2492 (P. 3), the author

wrote “Our recommendation is to use U+3FF00 through U+3FFFD for this new block, which provides 254 code points, and to use CJK Unified Ideographs Components as the block name to distinguish it from the CJK Unified Ideographs extension blocks.” As we know, U+2EBF0 through U+2F7FF in SIP have not been used yet, and there are 3,087 code points. Maybe we should use this part first, because it looks there will not be defined as any new CJK Unified Ideographs extension blocks.

For the block name, I suggest using “CJK Unified Ideographs Supplement”. I think this block could encode more types of characters which are used in the CJKV running text but it’s not better to include them as any CJKUI Ext block, such as [WS2021-00020:SAT-01301](#), [WS2021-00021:SAT-01303](#), [WS2021-00765:SAT-04332](#), [WS2021-00770:SAT-05240](#), [WS2021-00002:SAT-06315](#) and the atypical characters mentioned by me in [Section 4 of IRGN2413R2](#). The special case is [WS2021-00718:SAT-90136](#). It’s a ligature essentially, but this form is different from the original Siddham form showed by Maksim Sergeevich PERSIKOV, and we cannot get it by any method via the Han style of Siddham letters included in 《字孳補》. So, I think it’s OK to keep it in WS2021.

When I requested to encode the Gongche characters with my friends, WG2 tried to include them as the new block named “CJK Unified Ideographs Supplement” at U+2A6E0 through U+2A6FF, please see [WG2 N5006](#). As we know, those seven Gongche characters were moved to the end of CJKUI Ext. B finally, but I think we need to consider if it’s suitable to re-use this block name.

3. Preliminary proposal on the encoding method for the ideographic complex script(s)

As we know, the common Han script is not the complex script, but we have met some complex Han texts with the deepening of encoding works, although we know the text elements are based on the Han characters. It means all the “clusters” mentioned in this part are not suitable to encode in CJKUI in future directly.

3.1. Introduction on several situations

The ideographic complex script(s) here means some early Chinese organic chemical characters used in the paper *On the Nomenclature of Organic Chemistry* (《有機化學命名芻議》) written by Liang Kuo-chang (梁國常), the Sanskrit transcription and the Tibetan transcription used in the book 《同文韻統》, the Tangut transcription used in 《番漢合時掌中珠》 and the Jianzi Musical Notation.

3.1.1. Some early Chinese organic chemical characters

China NB submitted several early Chinese organic chemical characters used in the paper *On the Nomenclature of Organic Chemistry* written by Liang Kuo-chang to IRG WS2021 as below.

Table 3.1 Early Chinese organic chemical characters submitted to IRG WS2021

				
00016	00017	01900	00014	00777
GKJ-00941	GKJ-00942	GKJ-00943	GKJ-00944	GKJ-00877

Huang Junliang provided his comments under [WS2021-00014](#). I agree with him basically, but I need to show something different from him here.

In Liang's paper, he used two types of characters. The first one is shown above, one basic character with one numeral or one numerical sequence; the second one is the same as the common Han characters.

For the first type, we need to encode the following basic characters, and these basic characters are also needed in Liang's system.

Table 3.2 Basic characters of the first type of early organic chemical characters

1	2	3	4
𠄎	𠄎	𠄎	𠄎

And then, we also need one joiner. So, we could use the following sequences to represent the clusters.

<basic character, joiner, numeral>

<basic character, joiner, numeral, joiner, numeral>

In Huang Junliang's comment, he suggested using the glyphs for the above basic characters with dotted circle at the position of the numeral. When we check the original paper, we will know the ideographic glyph shapes of the basic characters as Table 3.2 shows are needed, and that will be suitable for CJKUI.

I suggest removing WS2021-00017:GKJ-00942, and changing the glyphs and the data for other China-Submitted characters in Table 3.1 as below.

Table 3.3 Updates for 4 China-Submitted characters

Current Glyph	𠄎	𠄎	𠄎	𠄎
Suggested Glyph	𠄎	𠄎	𠄎	𠄎
WS2021 SN	00016	01900	00014	00777
G-Source	GKJ-00941	GKJ-00943	GKJ-00944	GKJ-00877

The second type is the same as common CJKUIs, and three of them have been encoded in CJKUI, which are 𠄎 (U+2BB4D), 𠄎 (U+6C2C) and 𠄎 (U+930F), but the others of them have not been submitted by any submitters as below. I think all the characters shows in Table 3.4 are suitable to encode in CJKUI in future.

Table 3.4 Unencoded characters in the second type of early organic chemical characters

醞	醞	陽	錫	羸
脣	脛	體	旒	秬
饒	饒	饒		

Note that 𠄎身亞 has been submitted to IRG WS2021 by China NB as [WS2021-03927:GKL-00954](https://www.unicode.org/l2/WS2021-03927:GKL-00954), and the current evidence is questionable.

3.1.2. Sanskrit transcription and Tibetan transcription

These two systems here mean the ones defined in 《同文韻統》. This book defined four “levels” to transcribe the Sanskrit and Tibetan syllables to Han characters. The following table shows the examples for Sanskrit, and the Tibetan use is similar.

Table 3.5 Examples of 《同文韻統》

Type	Picture	Rationale	Siddham	Latin
1A		common CJKUI	𑖀	a
1B		fanqie character	𑖀𑖩	ña
2A		common CJKUI with a below small-sized common CJKUI which means the long vowel	𑖀𑖩	ā
2B		fanqie character with a below small-sized common CJKUI which means the long vowel	𑖀𑖩𑖩	nā

Type	Picture	Rationale	Siddham	Latin
3A		common CJKUI with a below small-sized common CJKUI which means the consonant	𑖆:	aḥ
3B		fanqie character with a below small-sized common CJKUI which means the consonant	𑖇:	ṅaḥ
4A		common CJKUI or fanqie character with a below small-sized common CJKUI which means the vowel is ɿ or ʅ	𑖈	kɿ
4B		common CJKUI or fanqie character with an above small-sized common CJKUI which means the consonant	𑖉	kra

For 1A, almost all the characters have been encoded, and some characters should be added to IVD. For 1B, I think the fanqie characters should be encoded as single Han characters in future. For 2A and 2B, we need one modified sign to represent the long vowel forms. For 3A and 3B, we need one sign like U+17D2 KHMER SIGN COENG. For 4A and 4B, we need a joiner; for 4B, the joiner here also means a virama. We could use one filler in 4A and 4B, which will be better for the encoding work and the education.

3.1.3. Tangut transcription

In 《番漢合時掌中珠》, the author use one Tangut transcription system to represent the Tangut pronunciations by Han characters.

The following table shows the different examples of this kind of Tangut transcription.

Table 3.6 Examples of 《番漢合時掌中珠》

Type	Picture	Tangut	Meaning	Note
1A		𑖇 U+17F3B	地	common CJKUI
1B		𑖉 U+17FF3	時	Tangut used CJKUI

Type	Picture	Tangut	Meaning	Note
2		𐰇 U+17E66	天	common CJKUI with the combing tone mark
3		𐰇 U+184D0	人	the right part is used to make the value of the initial different
4		𐰇 U+1735D	刑	the below small-sized part is used to make the value of the syllable different

For 1A and 1B, all the characters have been encoded in CJKUI.

For 2, it should be one CJKUI with U+302A through U+302D, which have been solved.

For 3, the use is similar to 4B in Table 3.5. The right part should be 尼 (U+5C3C), 魚 (U+9B5A), 泥 (U+6CE5), 你 (U+4F60), 溼 (U+57FF), 夷 (U+5937), 宜 (U+5B9C), 嘍 (U+20F2A) and so on. These characters are totally different from the characters used before virama mentioned in Table 3.5, so the sequence will not be confused. So, we need to use one joiner between two characters.

For 4, the shape is similar to 3A in Table 3.5. We also need to use one sign like U+17D2 KHMER SIGN COENG. The below small-sized part should be 合 (U+5408), 輕 (U+8F15), 重 (U+91CD) and so on.


3.1.4. Jianzi Musical Notation

In Jianzi Musical Notation, there are three types of typographic forms. The first one is called as 譜字 or 大字, the second one is called as 旁字 or 小字, the third one is called as 註字. The first one and the second one are necessary for all the Jianzi scores, but third one is not necessary, and it is similar to notes and commentaries beside one sentence by common or classical Chinese writing system, which should be handed by the typesetting software or the composition languages.

The following table shows the different examples of Jianzi.

Table 3.7 Examples for Jianzi

Type	Picture	Name	Note
1A		大指九徽勾四弦	
1B		大指七徽四弦	the main fingering letter has been same as before omitted

Type	Picture	Name	Note
2		上十徽八分	

In the Jianzi Musical Notation clusters, the main fingering letter is the most important for 1A and 1B. And we can distinguish them as different sub-types by the amount of the main fingering letter(s).

For 1A, we need to insert one joiner between two different fingering letters, markers, strings and numerals included.






For 1B, the main fingering letter is omitted. If we don't used one filler, it will be confused with others.


For 2, there is a glyph group of small-sized forms. And we need to use one small-sized form sign before only one CJKUI to represent the small-sized form, and the joiner(s) could make different small-sized forms become one cluster.

3.2. Proposal of the signs

I classified 5 signs as below based on the above analyses.

Table 3.8 List of the signs

SN	Type	Glyph	Name	Note
1	joiner		Ideographic Joiner (IDJ)	CJKUL,IDJ,CJKUI virama
2			Ideographic Long Vowel Sign (ILV)	CJKUI,ILV
3			Ideographic Auxiliary Below Sign (IAB)	IAB,CJKUI IDF,IAB,CJKUI CJKUL,IAB,CJKUI
4			Ideographic Small-Sized Form Sign (ISF)	ISF,CJKUI ISF,CJKUI,IDJ,ISF,CJKUI
5	filler		Ideographic Filler (IDF)	Same as CJKUI

SN	Type	Glyph	Name	Note
				

Note that the first one is very similar to zwj, but in the current use, the zwj is sometimes used for word recording and avoiding the single CJKUI at the beginning of one visual line (孤字成行). It is better to encode a new joiner only for CJKUI.

It looks IAB and ISF are similar, but they are different in fact. For the horizontal layout, there will be no any spacing for the basic character or sequence with the IAB sequence, but there could be spacing for the one with the ISF sequence. The <IAB,CJKUI> sequence should be treated as the combining mark. So, it's better to disunify them. On the other hand, I am considering if it's suitable to use ISF for the small *er* character discussed in [WG2 N4720](#) by Andrew West and me.

In East Asian Sanskrit studies, the joiner behavior is called as 合, so I suggest using 合 with dotted square for the alternate glyph for IDJ; the long vowel is called as 引, so I suggest using 引 with dotted square for the alternate glyph for ILV. The use of IAB could be called as 輔助 in Chinese, so I suggest using 輔 with dotted square for the alternate glyph. The use of ISF is called as 旁字 in Jianzi system, so I suggest using 旁 with dotted square for the alternate glyph. In Jianzi system, the fingering letters are sometimes written as 旨 which means the 指法, and in East Asian Sanskrit studies, the body is called as 體文, so there are two alternate glyphs, 旨 with dotted square could be used for Jianzi use, 體 with dotted square could be used for Sanskrit and Tibetan uses.

The following table shows the usages of these signs.

Table 3.9 Usages of the signs

Sign	chemistry	Sanskrit Tibetan	Tangut	Jianzi
IDJ	Y	Y (4A, 4B)	Y (3)	Y (1A)
ILV		Y (2A, 2B)		
IAB		Y (3A, 3B)	Y (4)	
ISF				Y (2)
IDF	Y	Y (4A, 4B)	Y	Y (1B)

4. Acknowledgements

Mr. Jerry You reviews this document.

Mr. Clerk Ma discusses the encoding method of Section 3 with me and confirms it's OK to run by OpenType with me.

(End of Document)