ISO/IEC JTC1/SC2/WG2/**IRGN2578**
2022-10-13

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

Doc Type:	**Ideographic Research Group Document**
Title:	**Comments on encoding early Chinese organic chemical character in WS2021 and other complex ideographs**
Source:	**Eiso Chan (陈永聪, Culture and Art Publishing House)**
Status:	**Individual Contribution to IRG #59, online meeting**
Action:	**For consideration by IRG and China NB**
Date:	**2022-10-13**

China NB submitted several early Chinese organic chemical characters used in the paper *On the Nomenclature of Organic Chemistry* written by Liang Kuo-chang (梁國常) to IRG WS2021 as below. Please see Table 1.

Table 1 Early Chinsese 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 on IRG ORT. I agree with him basically, but I need to show something different from him in this document.

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

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

Table 2 Unencoded characters in the second type of early organic chemical characters

Note that the corresponding simplified form ⿰身亚 of ⿰身亞 has been submitted to IRG WS2021 by China NB as WS2021-03927:GKJ-00954.

For the second type, we need to clarify what the complex ideograph is first.

Suzuki-San put forward a concept with the term “CJK complex ideographic symbol” in his April proposal WG2 N4796 in 2017. At that time, he tried to solve some problems about U+30EDD (邇) and U+30EDE (邇). I once discussed this issue with Suzuki-San and other experts, and my comment was it was not suitable to treat them as the complex ones, but there were the real complex ones in the real world. U+30EDD (邇) and U+30EDE (邇) have been kept in CJKUI at last, which was a correct decision. However, we need to meet the real complex ideograph (not symbol) in the ongoing and coming encoding works.

As we know, UCS and Unicode distinguish the scripts as the simple scripts and the complex scripts. So many Brahmi-related, Arabic-related scripts and so on are treated as the complex scripts, because they are the scripts where rendering aren’t just simply putting glyphs side by side, that include stuff with combining marks, ligatures, reordering, stacking and so on. CJKUI and the similar scripts like Yi, Tangut and so on belong to the simple script.

In general, CJKUI (even Tangut) should be distinguished as indivisible ideograph (獨體字) and compound ideograph (合體字). The indivisible ideograph includes only one component commonly, or it can’t be divided into more than one component; the compound ideograph includes more than one component or character. The components or the characters used as the components own their original readings, meanings and rationales, but the users don’t need to care about them when they read the compound ideographs in the running texts, which the compound ideographs own their stable readings, meanings and rationales, so they can be used the same as the indivisible ideographs as the individual head ideograph in the dictionary and the basic element in the running text. Therefore, the multi-component ideograph used for Chinese dialects, Zhuang, Vietnamese, Bouyei, Bai, Taoism and Buddhism all belong to the compound ideograph. On the other hand, there is no definite splicing rule of the activity for the rationales and typographies from multiple indivisible ideographs to the compound ideographs, and the compound result for the compound ideographs are not free. This is the reason why we still need to collect, review, design and encode the compound ideographs one by one up to now, instead of only encoding a bunch of indivisible ideograph and specific components. Even though the dynamic ideographic type design (动态组字) will be realized perfectly through AI possibly in future, the IRG works of encoding ideographs one by one still shall not be non-stop for a long time to come. We should call the indivisible and compound ideographs are the common or simple ideographs.

The complex ideographs are not different from the compound ideograph. The reading, meaning and rationale of complex ideograph come from the basic elements, and the futures of

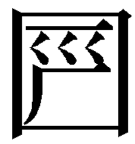
basic element and complex ideograph are closely related with each other. The splicing results for the complex ideographs are free if basic element and rule are stable for the sequences, which the splicing rule is systematic in one kind of complex ideograph. The users do need to care about the reading, meaning or rationales of the basic elements when they read the complex ideographs in the running texts. This situation is similar to the old Hangul syllables. On the other hand, I have finished the Jianzi encoding research, and we get more than two hundred million reasonable results and as least one million common ones used for Guqin.

Also note that the common qieshenzi (切身字) belongs to compound ideograph not complex ideograph because the compound rule and station for qieshenzi is not free and systematic and the meaning and reading for one qieshenzi is stable for everywhere. Qieshenzi means an ideograph compounded of two existing ideographs acting as an initial (聲母) and a final (韻母), sometimes with the semantic element. U+20193 (𠆓) is the error form of U+20199 (𠆙), but the reading is still not changed; U+2B9F6 (𫧶) is a qieshenzi only for Cantonese, but the real use is only limited to the word leng4keng4 (靈𫧶), other word read as keng4 in Cantonese can't use this character, and the cognate word in other Yue sub-dialect could also use this character despite the readings are not totally the same as Cantonese.

When we focus on WS2021-00016:GKJ-00941 (⿰⿳一巛⿸厂一充) and 00017:GKJ-00942 (⿰⿳一巛⿸厂二充), the main components for these characters are both ⿰⿳一巛厂充, and the positions of the residual numerals are stable, that means the splicing results of two specific elements are also stable without additional position information. WS2021-00016:GKJ-00941 (⿰⿳一巛⿸厂一充) means "methane" (甲烷 as current Chinese term), WS2021-00017:GKJ-00942 (⿰⿳一巛⿸厂二充) means "ethane" (乙烷 as current Chinese term), but the numerals could be used boundlessly as long as chemical technology is developed enough. All the results have been stipulated for the names, meanings and glyphs based on the rules defined by the author. Therefore, there is no need to limit only two results for the uses, because the basic part is free to be spliced with more than or equal to one numeral. Other China-submitted "characters" mentioned in Table 1 are similar to this pair.

Firstly, we need to encode the following basic characters, and these basic characters are also needed in Liang's system. I once mentioned this comment in my feedback on IRGN2551 as 4.6.

Table 3 Basic characters of the first type of early organic chemical characters

1	2	3	4
			

And we need to check if the numeral greater than 10 had been used in the historical uses. If not, the best way is to use 廿/卅/卌 with the basic numerals like Jianzi Musical Notation used for Se.

Secondly, we also need one joiner. I will show the detail of joiner later in this document.

Thirdly, we could use the following sequences to represent the clusters.

<basic character, joiner, numeral>

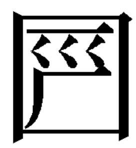
<basic character, joiner, numeral, joiner, numeral>

...

As I analyzed above, the glyph or glyph sequence for the final clusters results of the sequences are stable when the basic character and the numeral are stable. For example, the sequence <#1

mentioned in Table 3, joiner, U+4E00> has only one present glyph, so the font should provide the stable glyph(s) like WS2021-00016:GKJ-00941 (⿰⿱一巛⿸厂一充) not make the graphic element to left and right, top and bottom, or others, when the specific joiner is used in the sequence.
In Huang Junliang's comment, he suggested using the glyphs for the above basic characters with dotted circle at the position of numeral. When we check the original paper, we will know the ideographic glyph shapes of the basic characters as Table 3 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 1 as below.

Table 4 Updates for 4 China-submitted characters

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

For the joiner, I suggest encoding the new one separately not using ZWJ or CGJ, because ZWJ and CGJ are needed to use for other situations such as avoiding the single CJKUI at the beginning of one visual line (孤字成行) for text shaping, the horizontal parenthesized ideograph used in the vertical layout for the dulcimer performance of Yueju Opera (粤剧) and so on.
The name should be IDEOGRAPHIC JOINER and the abbreviation could be IDJ, because this joiner and the corresponding rule could be also used for other complex ideographs. The glyph of this joiner could be designed as below in the code charts.

Fig. 1 Glyph for Ideographic Joiner

In my ongoing research, the similar splicing rule could be used for more than 10 kinds of complex ideographs, so it is better to share this joiner for them correspondingly.
This is not a proposal for UTC, so I don't list the UCD information in this document now.
I will clarify the rules for the following kinds of complex ideographs in future, and it is expected that there will be five format controls and more than 50 rules. The following list is not complete. I am doing the research on two kinds of musical notation to record the Pipa performance for Quanzhou Nanyin and Chaozhou Xiyue, one kind of phonetic system in Ming and Qing Dynasty.

(1) Used for languages

Tongwen Yuntong used to record Mandarin in Qing dynasty, Sanskrit and Tibetan
Sanhe Qieyin used to record Manchu, Mongolian, Uigur, Kazakh and so on
Tangut-Chinese pronunciation indicating method
English-Cantonese pronunciation indicating method
Orthodox Russian-Chinese pronunciation indicating method
Tibetan-Chinese pronunciation indicating method

(2) Used for typography
typographical complex ideographs

(3) Used for traditional music
Jianzi Musical Notation used for Guqin
Jianzi Musical Notation used for Se
Gongche Musical Notation
Ersi Musical Notation
Hua's Pipa Musical Notation
Luogujing

(4) Used for early punctuations
Qianlong and Jiaqing Scholars used punctuation
Literal Musical Notation used punctuation

(5) Used numerals
Suzhou Numerals

Acknowledgements
Jerry You provides so many useful comments on the terms.
Clerk Ma helps me confirm the encoding method could run by OFF and OpenType.

(End of Document)