Doc Type: **Ideographic Rapporteur Group Document**
Title: **Proposal to improve IRG process**
Source: **Wang Xieyang (王谢杨)**
Status: **Individual Contribution**
Action: **For consideration by IRG**
Date: **2020-09-02**

Here are four parts of the document discussing the following issues:
1.Ways to deal with the characters wrongly put in D-set.
2.Avoid encoding Shengzaozi Characters(生造字).
3.Concretize the process of defining non-congnate characters.
4.The possibility of unifying some rarely used ideographs regardless of their meanings.
If the suggestions are confirmed by IRG, the IRG PnP should be changed accordingly.

1. Ways to deal with the characters wrongly put in D-set

In IRG#56, we found two Vietnam characters wrongly put in the D-set due to editorial error. After IRG#56, three China characters were found wrongly postponed:

https://hc.jsecs.org/irg/ws2017/app/?find=GXM-00267
https://hc.jsecs.org/irg/ws2017/app/?find=GXM-00303
https://hc.jsecs.org/irg/ws2017/app/?find=GXM-00370

New evidences of these three characters are posted and accepted in July, 2020(after the time of discussion record), but none of them are moved from D-set to M-set. I don't think we should blame this to anyone while we can absolutely make this happen less.
Here are my suggestions to deal with the problem:

In order to avoid this happening, I suggest we:

1.Skip no comment tagged *unification*, *evidence*, *glyph design & normalization* and *other* in our meeting.

2.In meeting mode, IRG ORT should give a special color to D-set characters who have new comments before or during the meeting.

3.We should get clear agreement from experts concerned or other experts who represent them that their new comments about the D-set character are well discussed in our meeting.

However, we can't assure that this or other kinds of errors won't happen again. So the messures to deal with it should be added to IRG PnP.

I suggest that we create a document for this kind of editorial errors every time we find it. If the M-set is not frozen, the characters wrongly put in D-set should be added back to M-set immediately; if the M-set is frozen already, the characters should be added to the next IRG working set WITHOUT being counted into the quota. The document work can be done by the editors of each source or someone designated.

2. Avoid encoding Shengzaozi Characters(生造字).

Encoding characters created by oneself in modern times can be dangerous. **Provided this kind of characters can be encoded freely, everyone who can afford to publish a book is able to creat a character himself and get it encoded.** That's also why, in practice, we have long been trying not to encode this kind of characters. Usually, we call these as Shengzaozi Characters(生造字).The defination of Shengzaozi Characters(生造字), however, can't be found in IRG PnP.

Considering some Shengzaozi Characters can be accepted by more and more people and then become Suzi Characters(俗字, also called Common Characters), the scope of Shengzaozi Characters is hard to defined. I suggest that we not define Shengzaozi Characters in IRG PnP but add one item to IRG PnP 2.1.1 said

> "**e. Scope of use（使用范围限制）**: The character should be used in running text by someone except the creator. It will be rejected if a character is considered created by one of the authors of the materials or reference which the proposed evidence is derived.

Since place or people name used characters can be used daily, the encoding of them won't be excluded.

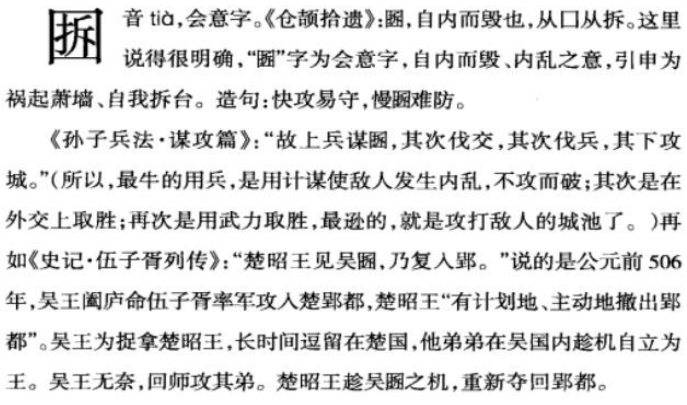
In IRG WS2021, UK and UTC submit some dialect used characters which are aparently created by the author. These characters are from《简明粤英词典》(杨明新 著, 广东高等教育出版社, 1999年) or *The Representation of Cantonese with Chinese Characters (Journal of Chinese Linguistics, Monograph Series Number 18, 2002)* without other use in running text. What's more, the actual source of UTC's evidences is also《简明粤英词典》.For example, UTC-00742:

The Representation of Cantonese with Chinese Characters, Journal of Chinese Linguistics, Monograph Series Number 18, 2002: p457, pos08

141 / 05	FA61 ud	𡃁	haa1	to bully; take advantage of so.	~霸	*~baa3* bully and humiliate; one who loves to bully others	YMX 1999:117

Fig.1 The evidence of UTC-00742 from IRG ORT

There is a book named 《造文字的反 一个草民的造字运动》(余少镭 著, 广州: 花城出版社, 2011 年 1 月) which contains many characters that the author creates. Every character has its own pronunciation, meaning and even running text.



圄 音 tià，会意字。《仓颉拾遗》：圄，自内而毁也，从口从拆。这里说得很明确，"圄"字为会意字，自内而毁、内乱之意，引申为祸起萧墙、自我拆台。造句：快攻易守，慢圄难防。

《孙子兵法·谋攻篇》："故上兵谋圄，其次伐交，其次伐兵，其下攻城。"（所以，最牛的用兵，是用计谋使敌人发生内乱，不攻而破；其次是在外交上取胜；再次是用武力取胜，最逊的，就是攻打敌人的城池了。）再如《史记·伍子胥列传》："楚昭王见吴圄，乃复入郢。"说的是公元前 506 年，吴王阖庐命伍子胥率军攻入楚郢都，楚昭王"有计划地、主动地撤出郢都"。吴王为捉拿楚昭王，长时间逗留在楚国，他弟弟在吴国内趁机自立为王。吴王无奈，回师攻其弟。楚昭王趁吴圄之机，重新夺回郢都。

Fig.2 余少镭: 造文字的反 一个草民的造字运动, 广州: 花城出版社, 2011 年 1 月, P7

Basically, the characters in 余少镭's book have no difference with 杨明新's in some way. Thus, characters from the two books shouldn't be encoded unless other reliable evidences can be found. We will be able to reject such characters according to IRG PnP directly if we add an item as above.

3. Concretize the process of defining non-congnate characters.

Here is a part of a chart in the book named《韩国汉文古文献异形字研究》（吕浩 著, 上海: 上海人民出版社, 2013 年 12 月）

主形字	韓國漢文古文獻異形字	敦煌文獻異形字
備	備、俻、俻、俻、脩、倏、備、備、備	俻、備、備
邊	邊、辺、辺、邊、邊、邊	邊、邊、邉、邊
稱	稱、稱、稱、稱、稱、稱、稱、稱、稱、稱、稱	稱、稱、稱、稱
處	處、處、處、處、處、處、処、処	處、處、処、処
辭	辞、辭、辭、辭、辞、辭、辭、辭	辭、辞、辞、辞
帶	帶、帯、帶、帶、帶、帶	帶
殿	殿、殿、殿、殿、殿、殿、殿、殿、殿	墅、殿、殿、墅
發	発、發、發、発、發、發、發、發、発、發、發、發、發、発	發、發、發、發、發、發、发、发
歸	帰、歸、歸、歸、歸、歸、歸、歸、歸、帰、歸、歸、歸、埽、歸、歸、歸	帰、歸、歸、歸、歸、歸、歸、歸、歸、歸、帰、飯
龜	龜、龜、龜、龜、龜、龜、龟、龜、黿、龜、龜、龜	龜、龜、龜、龜
號	號、號、號、號、號、號、號、号、號、號、號	號、號、號、號、號、號、号、号、号、号

Fig.3 吕浩: 韩国汉文古文献异形字研究, 上海: 上海人民出版社, 2013 年 12 月, P27

This chart is just a tip of the iceberg. Many Han characters have a wide range of glyphs. It is very likely that one of the glyphs inherit one rare meaning of the character while other glyphs don't. Then even if the glyph has little difference from others, it is of great possibility that we disunify them. Actually, however, the users even don't care. Some characters are usually written as very similar glyphs in the practice and most people even don't think there is difference between them. In this case, the disunification is not very necessary.

In IRG PnP 2.1.3, it says

> Ideographs with different glyph shapes that are unrelated in historical derivation (non-cognate characters) are not unified no matter how similar their glyph shapes may be.
>
> ……
>
> Because shape analysis alone may not tell non-cognateness or semantic differences, it is the submitter's responsibility to provide information and supporting evidence in order to invoke the non-cognate rule.

It doesn't say, however, which kind of information and evidence is needed in spercific. What's more, IRG PnP doesn't clarify how to deal with the mixing used simlar glyphs. Consideing the large quantity of Han characters' glyphs and the conclusion of our discussion of GDM-00241(⿰木匋) in IRG WS2021 and UK-10757(⿰门丁) in IRG WS2017, I suggest

> 1) Add one sentence to the quoted paragraph 1 that "Ideographs with unifiable glyph shapes should be considered congnate if they can be used without distinction in fact".
>
> 2) Add one sentence to the quoted paragraph 2 that "For unifiable ideographs, the information and supporting evidence provied by submitters should be able to clearly explain a) the pronunciation of the two ideographs have no historical derivation; b) the meaning of the two ideographs have no relationship.

4.The possibility of unifying some rarely used ideographs regardless of their meanings.

After trying to concretize the process of defining non-congnate characters, I think there is a possibility to unify some rarely used **unifiable** ideographs regardless of their meanings. We can take it as two or more meanings of one character but two non-congnate characters if their meanings are unrelated. If the user does care for the shape, fixing the problem on the level of fonts or using IVS should be preferred then.

This is just a thought after brain storming, it will be very dangerous if we actually do it.

(End of Doc)