Title: IRGN2306 Proposal to Streamline IRG Processes

Type: Individual Contribution

Source: Henry Chan

Date: 2018/05/16

1. Response to IRGN2305 CJK Regional Supplementary Ideographs

1A)

The proposal in IRGN2305 is for IRG to cease conducting unification review and not to unify across regions.

Ceasing to unify across regions results in de-facto future disunification of CJKV Hanzi/Kanji/Hanja/Chữ Hán. The unification of CJKV Hanzi/Kanji/Hanja/Chữ Hán as a single unified script is a long-settled issue both in Unicode and mainstream linguists. This decision should not be overturned out of simple “efficiency” concerns.

Unification review is the main job of IRG. If IRG is relieved of the duty to conduct unification rule, and that IRG only checks that the evidence matches the glyph, we might as well dissolve IRG, and leave such secretarial works to other bodies, similar to non-CJKV scripts.

Having identical glyphs will create opportunities for spoofing and phishing, and harm the information exchange between encoded old manuscripts across regions. If CJKV is still considered a single script, WG2 will still have to conduct the unification review. I doubt it would be more efficient process for the individual ideographs to be discussed and argued in WG2.

1B)

The proposal claims that the Japan NB has been very concerned about the proliferation of variant characters and rarely used ancient/academic characters in CJK Unified Ideographs extensions.

Instead of proposing ways to adequately handle the archival of ancient/academic characters, Japan NB in the past meetings chose to primarily pick on derived simplified ideographs, many of which are actively in-use by China.

A few experts in IRG also insist on requesting unification for characters even when proof of non-cognate is already provided. In cases where the relationship between the shape difference and non-cognate is trivial, such suggestions for unification should not be made. IRG has clarified multiple times that UCV is a general guideline. Strict application of the UCV is counterintuitive and only wastes time.

UCS is not a dumping ground for historical glyphs and characters. To speed up the ideograph encoding process, the better solution is to provide suggestions how to improve the model and guidelines to pre-screen variants and improve the PnP to remove bottlenecks.

By clearing the unification review, that effectively turns on the green light for member bodies and members to encode their collections of obscure ideographs. The result is UCS ends up as a dumping site for any kind of character. IRGN2305 proposal actively acts against an orderly and restrained encoding of extension ideographs.

1C)

The proposal claims that most encoded glyphs are for a certain region and there is little sharing across regions.

The calculation in IRGN2305's calculation method is flawed. It does not take into account the pending horizontal extensions lined up for IRG. It does not take into account the horizontal extensions that are possible after the processing of the Working Sets. It does not take into account that some member bodies are reluctant to submit horizontal extensions as it gives more work than benefit to these countries.

Glyph sharing between regions is still expected in the near future. The Tày characters submitted by Vietnam are shared by the Zhuang characters in use in China. Considerable glyph sharing is foreseen.

Glyph sharing can increase if the unification is suitably relaxed. As seen in WS2017 reviews, many ancient Hanzi are under study by both Chinese and Western scholars, and they may pick components with different shapes with transliterating old characters (such as 攵 vs 攴). Had the model been effectively relaxed in the past, the single-sourced percentage would not be so low.

Currently, member bodies need to submit an extra document requesting horizontal extension after a working set, even if the member body had already indicated so during the working set review and the meeting decision is unanimous. This process could be adjusted so the percentage of horizontal extension accurately reflects the real condition.

The claim for unshared ideographs and the concern of proliferation of variants does not match the actual reality. One of the worst offender of proliferation of variants encoded thousands of vulgar variants mostly sourced from Longkan and other unconventional dictionaries into Extension F. Instead of establishing a more suitable encoding model, the submitter simply dumped all the characters without considering the systematic writing differences that resulted in these variants and was angry at the UK NB for having pointed out so.

The proposal is also released at an immature time because it has not given a chance for IRG to operate under the relaxed unification rules for WS2017 and determine which extent the encoding of variants can be prevented.

The main issue facing IRG now is that the sole evidence for a given character suggests that it is a rare historical form of another actively used character, but submitters disagree with unification because “oh, it looks too different”.

IRG is wasting time on discussing unification because the model is insufficient for member bodies to pre-screen their characters to be encoded via alternative means. The unification model is born out of inter-region preferences and is insufficient to cover the archaic writing orthographies.

For archaic characters, IRG member bodies are forced to conduct the unification rather arbitrarily because of the lack of guidelines and precedent to handle vulgar variants and error forms.

The solution is to evolve the model and set guidelines which take the expected variability of different sources and mediums into account.

“It’s not my block, so I don’t care” is ***ignoring*** the problem, not fixing the problem.

Allowing each region to have their own set free from unification review effectively strips other member body’s experts’ ability to review other region’s character sets. This will weaken IRG’s ability to discover and propose encoding models more suited to encoding different classes of variants.

Lastly, some national bodies have the tendency to claim “It is the same character, but the glyph shape looks too different” when faced with review to their character set, and “these two character looks alike, and they should be unified even if non-cognate” to other submitter’s. Such attitude is ignorant to the nature of the script and should be strongly refrained from. Just because “i” and “j” look so similar does not mean they are the same character; just as “a” and “ɑ” look so different does not mean they are not the same.

2. Concern to TCA's and SAT's large number of variants

2A)

A sizeable majority of glyphs submitted by TCA are from old inscriptions from bronze and stone. If IRG accepts the MOE dictionary as an authoritative source, IRG should accept the mapping of the variants to the character head as authoritative and approve such characters for unification via IVD/IVS technology automatically.

2B)

A sizeable majority of glyphs of TCA and SAT are vulgar variants coming from ancient Buddhist texts, 龍龕手鑑 and 四聲篇海 which is well known for including characters in an unconventional style. Character unification should be executed based on the full range of differences in vulgar variants in manuscripts at that time.

The current model is completely unsustainable; it will take at least 20+ years for TCA to finish encoding the characters inside the MOE dictionary. Three solutions are identified:

Solution (A)
IRG PnP should be amended such that member bodies are required to draft new UCVs that cover such cases. Member bodies would highlight the representative glyph forms and compile a list of variant glyph forms. This is a similar method to the normalization procedure carried out by ROK.

Solution (B)
IRG establishes a sub-group "Vulgar Variants Working Group" which is in charge of assessing the historical vulgar variants in 龍龕手鑑, 四聲篇海 and other manuscripts, then compiling a list of representative glyph forms and unifiable glyph forms. Characters which are in the unifiable glyph forms should be unified to the representative glyph forms via IVD/IVS.

Solution (C)
The proposal in IRGN2305: each region submits a separate proposal for encoding and approved by WG2 separately. This would drastically reduce the workload of IRG but would extremely hinder the efforts of culture preservation and information exchange among scholars studying old manuscripts. Since the vast majority of characters for TCA are vulgar variants, and TCA's encoding of conventionally-composed characters would easily be jeopardized by any WG2 opposition to encoding vulgar variants as is.

3. Improving accuracy of IDS check

After the IDS editor conducts the machine IDS checking, an expert should be appointed to screen the IDS and separate the list into (1) exact matches, (2) similar matches and (3) non-cognate characters according to submitter's evidence.

Characters already identified as non-cognate via the initial evidence provided by the submitters should not be consolidated into the comments nor discussed in the meeting unless there is other evidence indicating otherwise.

4. Improving process of comment submission & accuracy of discussion record

IRG review of characters is based on a two-step process with review and response to review. However, due to various reasons, member bodies and individual contributors may not be able to submit them in time before the deadline. Also, misunderstanding often occurs and may not be successfully resolved even after the meeting. The Chief Editor also needs to conduct the tedious task of manually consolidating comments by copy and pasting the word file. It is inevitable that some errors and/or omissions would easily occur.

The discussion record is recorded on the consolidated comments and other submissions and then consolidated onto the working set afterwards manually by the Chief Editor. Multiple experts have kept track of the unification record during the meeting and found that the recorded decision often differed.

<u>Suggestions</u>

It is suggested that an online review platform be instated for IRG such that member bodies and individual experts may directly leave their comments per ideograph, and the online review tool allows directly exporting to consolidated file.

It is suggested the meeting decisions will be directly recorded on the online review tool during the meeting allowing immediate confirmation by all attendees. The tool should also allow the export of the new version of the working set immediately after the meeting.

5. Stranding of postponed characters

If a character is postponed at a late stage of the working set preparation, or member body forgot to prepare a response, or IRG forgot to review the response, then a character is often left stranded in the last working set (e.g. WS2015) after the coming working set has already commenced (e.g. WS2017). The character would then unfortunately be postponed to the next working set (e.g. WS2019) and/or later.

Since submissions from a member body may in fact come from different departments, and such experts may miss some meetings, it has occurred that IRG often mistakenly misses some documents.

Suggestions

A rolling process could be adopted. First, submitters submit to a Central Registry by IRG as “A-set”. The characters will be reviewed for any unification possibilities and attribute problems. The submitters may directly accept any proposed changes without needing to pass through the meeting.

Before each meeting, submitters will nominate a set of characters to progress into a new “B-set” from “A-set” based on the quota assigned by the previous meeting. Only characters that have been submitted to the Central Registry for at least 3 months may be nominated. If any issues are unresolved in the meeting, the character is moved back into the “A-set”. Any characters that progress into the “B-set” cannot be modified unless decided in a meeting.

Before each meeting, member bodies and individual experts will review the current “B-set” ideographs to identify unification and attribute issues. If any issues are unresolved in the meeting, or any character is under suspicion, the character is moved back into the “A-set”, otherwise moved into the “C-set”.

The “C-set” is then submitted directly for WG2 approval. Individual characters may only be removed from the “C-set”.

If a character is resolved to be candidate for horizontal extension in B-set or C-set, it is moved to the H-set and submitted directly for WG2 approval. If issues are found, it is moved back to the A-set.

If a character is resolved to be candidate for IVD in B-set or C-set, the submitter may directly apply to UTC for IVD registration without further confirmation from IRG.

If a character is resolved to be unified to a certain character, and the existing glyph shape to be changed, the glyph change / glyph identifier change request may be submitted directly for WG2 approval without further confirmation from IRG.