Universal Character Set

UCS

ISO/IEC JTC1/SC2/WG2 IRG N 2216

Date: 2017-06-16

Source:	Japan Experts
Title:	Proposal regarding to the development of WS2015 and further extension of CJK Unified Ideographs, and Improvement of IRG's working process
Distribution:	IRG #48
Medium :	Electronic

[Description]

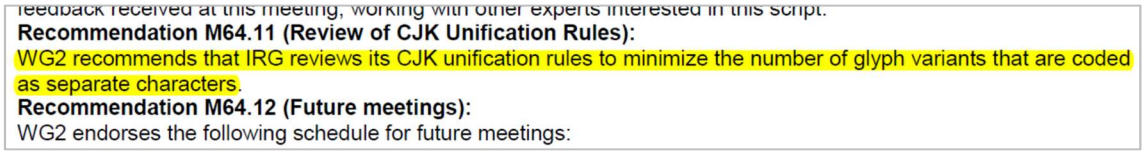
Although ISO/IEC10646 already contains more than 87,000 CJK Unified Ideographs including CJK extension F, more characters are still requested to be added. Actually WS2015 contains more than 5,200 candidates. According to the analysis result of the proposal for these candidates by Japanese experts, it was found they can be roughly categorized as follows:

(a) Variations of already encoded Unified Ideographs that cannot be unified due to the unification rules.

(b) Rare characters that are appeared only in ancient or academic publications (including dictionaries), and not used generally. Derived simplified is a typical case.

Recent IT systems are getting so powerful that they can handle many characters. However, proliferation of characters without careful consideration on variations will require a big cost of processing on IT systems such as editing or searching information as well as bigger storage space for dictionary and fonts.

In addition, at SC2/WG2 #64 meeting in 2015, WG2 recommended IRG to review CJK unification rules to minimize the number of glyph variants as separate characters.


feedback received at this meeting, working with other experts interested in this script.
Recommendation M64.11 (Review of CJK Unification Rules):
WG2 recommends that IRG reviews its CJK unification rules to minimize the number of glyph variants that are coded as separate characters.
Recommendation M64.12 (Future meetings):
WG2 endorses the following schedule for future meetings:

Following this recommendation, Japan experts request IRG to discuss:

1. Revision of the current unification rule to stop adding many variations to the existing

ideographs. When it is needed to handle such variations, use alternate method like IVS (Ideographic variation sequence). See appendix A for details.

2. Characters that are unsure to be used in real such as derived simplified, should be handled separately from CJK Unified Ideographs. For example, such character should be encoded in another block or use IVS as a variants. See appendix B for details.

Outcome of the discussion should be applied to the work of WS2015 to confirm its effectiveness.

Japan experts also propose followings which are derived from the past review work on WS2015.

- If proposing characters of which evidences are only in handwritten or wood block documents, the submitter should ensure in advance their validation of the normalization so that IRG can avoid discussion about the correctness on the submitted glyphs.
- To make IRG's review on CJK Unified Ideographs efficient, machine checking method should be used actively to find duplications or error attributes in addition to eye-ball review.

See appendix C for details.

Appendix A. Re-evaluation of Unification rule

1. Enlarge the range of unification regarding the differences of component shapes

[Background]

The current unification rule described in Annex S, was firstly drafted following the development of CJK main block (aka URO) and established while developing CJK extension A. This focuses popular and stable shapes that are well implemented and used as regional standards for many years, however it is not effective on the recent submissions because so many unfamiliar shapes are included. Although IRG has been making effort developing IWDS to compliment unification rule, it is not good enough.
The table below shows how the characters in CJK ideograph blocks are efficiently unified. It can be concluded that late collections consist so many single source characters.

	#source	#code point	difference	#sources per code point
URO	102,426	20,983	81,443	4.88
Ext. A	18,753	6,582	12,171	2.85
Ext. B	72,925	42,711	30,214	1.71
Ext. C	4,532	4,149	383	1.09
Ext. D	226	222	4	1.02
Ext. E	5,790	5,762	28	1.00
Ext. F	7,649	7,473	176	1.02
total	212,301	87,882		2.42

Summarize from CJK.txt of DIS 5th ed.

According to the submitted evidences, we can see that there are many variants of already encoded characters, which are not "new" characters. In other words, several code points are assigned into "one" character. Such assignment is compliant to the existing unification rule. To avoid this situation, it is necessary to revising the unification rule.

[Proposal]

Start discussion regarding unifiable shape differences based on the discussion of the past review. Apply the conclusion to the work of WS2015. It should not be applied to the already adopted collections including extension F to avoid confusion.

Most of these characters are not practical so their addition is basically a great cost of the process such as searching or editing electronic documents not only the space consuming

of dictionary or fonts. So it is reasonable to restrict adding more variations. For this purpose, it is necessary to revise the rules. Variations should be handled using more appropriate method like IVS.

2. Restrict using non cognate rule

[Background]
Non cognate rule is sometimes used vaguely. It is often used by the reason of the difference of meaning or pronunciation without explanation about cognate in the authentic dictionary. It is quite natural that two (or more) cognate characters have different meanings or pronunciations because each of them has been established in different regions with its culture and customs for long time.

[Proposal]
Non cognate rule should be restricted using only when characters can be explained their cognate in the authentic dictionary.

Appendix B. Proposal about handling characters those are unsure used in real world

[Background]

IRG PnP prohibits handling derived simplified characters if actual usage is not submitted. In other words, characters only appeared in specific dictionaries is out of IRG's scope. On the other hand, IRG PnP also describes about handling characters not appeared in any dictionary, such as person's name. Japan experts would like to reconfirm this policy first.

On the other hand, Japan experts also recognize that some people would like to use those characters as UCS..

[Proposal]

IRG should consider the solution how to handle unsure characters (including derived simplified) separate from CJK Unified Ideographs. The method(s) on this issue should seek approval of WG2. Possible solutions might be:

- Encoding such characters into different block from CJK Unified Ideographs. In this case, it is needed to request WG2 to establish new block for them.
- Derived simplified characters should be expressed as IVS (Ideographic Variation Sequences) to corresponding traditional characters. This solution may request revision of UTS37 and such derived simplified characters are need to be registered to IVD.

It is also necessary to discuss about derived simplified characters in particular because it can be created algorithmically.

Appendix C Improve IRG's working procedure

1. For characters from cursive sources such as handwritten documents or woodblock prints, submitter should be responsible on their correctness and stability of normalization.

[Background]
In Annex S of ISO/IEC 10646, unification procedure is based on Song/Ming style glyph shapes. In accordance with this Annex, IRG PnP requests submitter(s) to prepare Song/Ming style glyphs. Recently it is sometimes questioned that the correctness of the glyph shapes because they have only cursive evidences such as handwritten documents or woodblock prints, and many characters have been modified their glyphs after discussion. This kind of issue should be resolved before submission to WG2. During CJK extension F review, ROK submitted normalization rule (IRGN2154) to make clear their idea. Such document is reasonable avoiding waste of discussions.

[Proposal]
If characters from cursive sources are proposed, submitter should ensure the correctness and stability of their glyphs in advance. For example:

(1) Official document(s) regarding normalization are associated when proposing such characters, and submitter will maintain the document(s) continually. The submitter should respond appropriately in case the questions on their rule are raised.
(2) Information about submitted characters including usage will be opened and can be accessible online in a way maintained by the authentic third party such as:
   - Government, or e-government related initiatives (cf. 2.2.3a of IRG PnP)
   - Academic organization
   - Standardization organization
   - Consumer group

Endorsed procedure should be discussed at IRG.

2. Using machine checking actively in addition to eye-ball reviewing

[Background]
As mentioned, the number of coded CJK Unified Ideographs are more than 87,000 and it is very difficult to review all proposed characters perfectly to avoid duplications. IRG already use IDS (Ideographic Description Sequences) database to list possible

duplications efficiently, it is successful to reduce overlooking possible duplications. However, the situation is always changed according to the growth of the number of characters, progression of IT technology and capacity and dependence on the international standard as a basic infrastructure. IRG is expected to maintain and manage its electronic resources such as bitmap, fonts, attributes and discussion records so that they can be licensed to the third party that will develop various systems to utilize UCS under the appropriate conditions.

Although IRG sometimes discussed and reconfirmed on the conditions of using electronic resources, it is necessary to continue discussing in accordance with the change of the situation.. In Japan, for example, there were a complex license conditions for fonts used for extension E and former collections. However, due to the release of single alternate font with reasonable license conditions by Moji-Joho kiban project, it becomes much easier for everyone to use J-source glyphs adopting in the systems.

[Proposal]

- IRG should invite its members or related parties to demonstrate their systems if they can be shared with other IRG members for review purpose. If it is effective to reduce working cost, IRG will encourage its members to use the system (voluntary or mandatory)
- IRG should maintain and manage electronic resources regarding existing and proposed characters such as bitmaps, fonts, attributes, discussion records and so on, so that IRG can license them to anyone who intends to develop machine checking systems smoothly.