Universal Multiple-Octet Coded Character Set
UCS

ISO/IEC JTC1/SC2/WG2 IRG N1310
Date: 2007-05-31

Source:	John Knightley
Title:	Necessary and Sufficient Evidence
Status:	For discussion
Distribution:	IRG Members and Ideographic Experts
Medium:	Electronic

In deciding whether or not two characters should be unified the two most common challenges are (1) to reach agreement on whether or not the two characters have the same abstract shape and (2) to get sufficient information to decide whether or not the characters are cognate. In this document the second challenge is discussed.

Though a lot has been said about looking at the abstract shape of characters, abstract shape is not mentioned until S.1.2, before that is S.1.1

> 'S.1 Unification procedure
> S.1.1 Scope of unification
> Ideographs that are unrelated in historical derivation (non-cognate characters) have not been unified.'

To ignore the reading and meaning of characters is a recipe for eventual disaster. Though showing that a new character has a different abstract shape to all encoded characters is all that is necessary to prove that it should have its own codepoint, however this is not sufficient for the long term work of the IRG where the cognate issue must be addressed. Priority should be given to those characters with sufficient supporting information and more supporting information should be added to the remaining characters.

87% of the characters already encoded come from at least one of four widely distributed dictionaries, either Kangxi Dictionary, Daikanwa Jiten, Hanyu Dazidian or Daejaweon. The characters in these have already been encoded. Further Unihan.txt contains a dictionary reference of all but 8462 of these.

kIRG_TSource 54992 entries 6874 (12.5%) no dictionary reference
kIRG_GSource 57616 entries 1477 (2.5%) no dictionary reference
kIRG_KSource 1766 entries 111 (6.3%) no dictionary reference
kIRG_JSource 13178 entries 0 (0%) no dictionary reference
(based on figures posted by R Cook on the Unihan mailing list)

Starting with extension C most new characters either do not come from a dictionary, or come from less well known dictionaries. The IRG now correctly requires evidence for each character submitted however the exact form and content of this evidence is largely decided by the submitter, some submissions go into great detail but others are very spartan.

When checking extension C for duplicates, candidates where either the submitted character or the encoded character have neither reading, meaning or dictionary reference, take many times the effort per case to resolve than the rest. Rather than encoding more characters without sufficient evidence to decide if they are cognate or so making processing characters harder, in fact a concerted effort should be made to document well those encoded characters that lack such information at present.

Great attention has been paid over the years to standardizing information relating to the shape of characters, such as radical, first stroke, stroke count, and IDS. Attention also needs to be paid to the evidence given, particularly that required to decide whether or not a pair of characters are cognate. Though the details need to be fleshed out several seem obvious:-

(1) Printed references (Vol-page-line/entry) for those which come from dictionaries.

(2) Reading — the pronunciation

(3) Meaning — some sort of gloss

There are good reasons to do this:-

(1) The inclusion of such evidence will in the long term speed up the work of the IRG and ensure the quality of it's work. Though once encoded separately it is impossible to unify two characters, but two unified source glyphs can be disunified IRGN956_Unification.pdf states:-

"Because of this, the recommendation of the Ad Hoc group is that where there is doubt as to whether two forms should be unified or not, the null hypothesis is in favor of unification—that is, the editors should recommend unification in the absence of evidence to the contrary."

It should be noted that being non-cognate is the only ground for disunification. "Disunification request with reason of mis-application (over-application usually) of unification rule should NOT be accepted due to the principle in resolution M41.11." (WG2 n3102.doc, page 57)

The cost of any mistake is great and with thousands of characters to process, in practice the best solution may in fact be to move back doubtful characters to a later extension.

(2) The inclusion of such evidence would give credibility to the collections of characters submitted by the IRG to WG2. If several thousand characters are added to Unicode an no one knows what they mean, then some will complain that confusion is being caused. If, by comparison, several thousand characters are added and the reading and meaning are easily available then many will say how interesting and wonderful it is.

(3) The present mechanisms, both human and digital, for deciding on the abstract shape of characters is still being perfected. One major part of the learning process is to be able to assess different ideas about abstract shape, for this accurate readings and meanings are essential.

Though adding the reference, reading and meaning to the evidence for a character seems to be extra work in the long term it will save both time and energy. The inclusion of both reading, meaning and dictionary reference is already a trend, for instance IRGN1305 includes all three. Since these three things are so important, agreement should be sort upon a standard format suitable for inclusion in a database.