

ISO/IEC 10646/JTC 1/SC 2/WG 2/IRG  
**Ideographic Research Group (IRG)**

**Issues on Transliteration of Ancient Scripts**

Response by John Knightley – part 2(2023-15-10)

The vast majority of characters submitted to the IRG whilst reviewed are rightly processed with little or no discussion because the data is clear and only contains, and a minority require extended discussion either in writing or verbally in meetings. Members therefore are encouraged to make high quality submissions that can be handled within the time frame allowed. If too many characters require extended discussion then problems can occur. It is suggest the IRG make to PnP the changes below to improve the quality of submissions and efficiency of processing working sets. These suggestions falls into two parts, the first part, for the IRG, to explicitly increase the scope of checking of preliminary submissions by the IRG, the second part, for submitters, a requirement to limit the number of long or ambiguous IDS in a submission.

1) Increasing the scope of IRG checking of preliminary submissions

Currently according to PnP at the preliminary stage the criteria considered in whether or not the IRG should request a member to reduce the size of their submission is if the number of characters is too great. Increasing the scope of checking at the preliminary stage would allow some problems to be dealt with by making changes between the provisional and final submissions, including in some situations requesting a submitter to reduce a submissions. This could be done by adding a extra section(s) to PnP 3.1 *Call for Submission* at the end or elsewhere. The wording could contain specific targets such as:

- f. If a review of a submission suggests that extended discussion might be required for more than 10% of a submission of more than 500 characters or more than 50 characters of a submission less than 500 characters then the IRG may request the submission be reduced in size.

Here 10% is based firstly the on the estimate that if there is extended discussion of a character there is a 50% chance of it being unified so such a submission is likely to break the 5% rule, and on the other based on an estimate on the time available for discussion in IRG meetings and the fact the the discussion time for most characters is no longer than 2 or 3 minutes.

Or could be more general in wording , such as:

- f. If a review of a preliminary submission suggests that it may prove difficult to process in the normal time frame of a working the IRG may request the submitter to change or reduce the size of a submission.

These are not mutually exclusive, so if desired both be added.

2) Submitters limiting the number of long or ambiguous IDS in a submission

Whilst this might at first seem difficult to limit long or ambiguous IDS in a submission, it should be noted, the majority of members already make submissions that, for whatever reason, fall well within the limits suggested below. Many extended discussions of characters rightly center around the question of unification and since it is expected that the number of unifications be low (cf the 5%

rule) the time available in IRG meetings means that the number of characters in a working set requiring extended discussion about unification must also be quite low for a working set to progress smoothly. Ws2021 figures suggest that well under 10% of characters are discussed for more than 5 minutes in meetings. One strong indicator that the the unification of characters being more likely to require extended discussion is if the percentage of long or ambiguous IDS in a submission is high. Given a clear definition, short and unambiguous IDS in a submission can be easily calculated, everything else is a long or ambiguous IDS. Possible wording, that could be added to section 3.1 e or elsewhere would be:

The number of long or ambiguous IDS a any submission for a new collection should not exceed the greater of 10% of the number of characters in the submission or 50.

Two definitions should also be added:

**Long or unambiguous IDS:** Any IDS sequence that is not a short or unambiguous IDS including empty, incomplete or malformed IDS sequence.

and either

**Short and unambiguous IDS:** An IDS consisting of one IDC, either 𐀀, 𐀁, 𐀂, 𐀃, 𐀄, 𐀅, 𐀆, 𐀇 or 𐀈 and followed by two IRG approved character description components.

or

**Short and unambiguous IDS:** An IDS consisting of one IDC, either 𐀀, 𐀁, 𐀂, 𐀃, 𐀄, 𐀅, 𐀆, 𐀇 or 𐀈 and followed by two IRG approved character description components, or 𐀉 𐀊 𐀋 with an IRG approved character description component in the middle.

Please note that whilst sequences like &S8-01; are IRG approved and so count as a single character description components, self designated sequences do not count as a single character description component.

To help members better evaluate the proposed changes a comparison to the ws2021 data at time of submission is shown below.

Table 1: ws2021 Comparison

Submission	Long or ambiguous	Percentage long or ambiguous (v1)	Change if 𐀉.𐀊 also short	Percentage long or ambiguous (v2)	
IRGN2483	93/1160	8.3%	-28	5.7%	< 10%
IRGN2484	12/191	6.3%	-1	5.8%	< 10%
IRGN2485	113/383	29.5%	-1	28.7%	
IRGN2486	153/1000	15.3%	-9	14.4%	
IRGN2487	89/1000	8.9%	-4	8.5%	< 10%
IRGN2488	20/153	13.1%		13.1%	< 50
IRGN2489	51/1001	5.1%		5.1%	< 10%