

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

Issues on Transliteration of Ancient Scripts

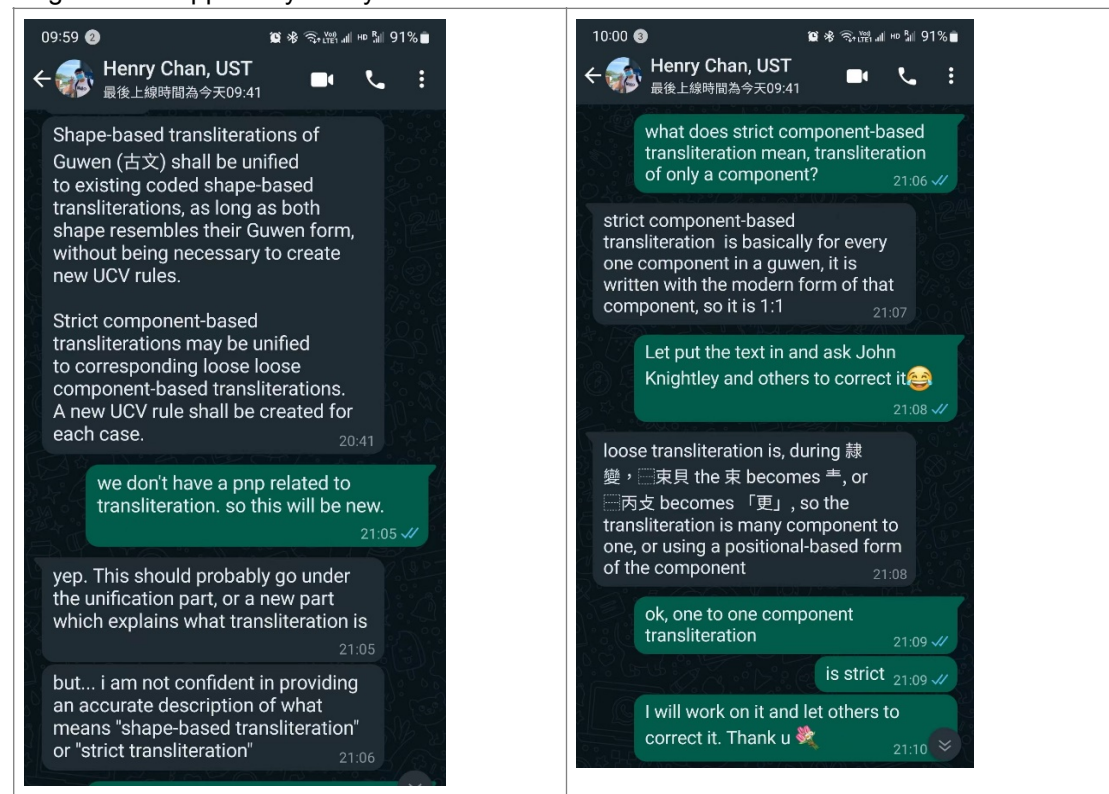
During the discussion in IRG WS 2021 V4.0 in IRG 60, IRG considers it important to address the issue of transliteration of ancient scripts. As the issue is complex and there are different transliteration definitions and works, IRG experts are asked to provide more information on transliteration. Below is a draft text intended as 2.1.5 of IRG PnP. The original text from Henry Chan is included for reference.

Draft text for IRG PnP on the handling of transliteration:

2.1.5. Handling of Transliterations of Guwen(古文)

Glyph shape-based transliterations of Guwen shall be unified to existing coded shape-based transliterated characters as long as both shapes resemble their Guwen form. In this case, there is no need to create new UCV rules/examples. 1-to-1 component based transliteration may be unified to its corresponding loose transliteration which may not have a 1-to-1 mapping of all the components. In this case, a new UCV rule should be created for each of such cases.

Original text supplied by Henry Chan:



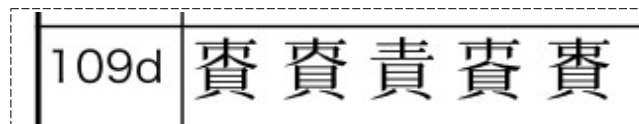
ISO/IEC 10646/JTC 1/SC 2/WG 2/IRG
Ideographic Research Group (IRG)

Issues on Transliteration of Ancient Scripts

Response by John Knightley – part 1(2023-10-10)

Whilst at first sight it may seem that the way best of dealing with characters formed by transposition from more ancient scripts and their many transitional variants, to do so would be to have a different set of unification rules for a subset of characters, and this would be inconsistent with the approach already used for over 30 years since the end of the source separation rule in which all characters are treated the same regarding unification, it would require ignoring the abstract shape of those character and therefore as such does not appear suitable for inclusion in IRG PnP.

The proposed addition to PnP leans heavily on the ideas behind the unification of 責 and 𠂔束貝 in UCV #190d, however a re-examination will show that this decision was inconsistent with UCS precedent and that 責 should be removed from UCV 190d.



(Figure 1: UCV #190d)

This addition was made based on the unification of SAT-06454 𠂔石束貝 and U+78E7 𠂔 (see <https://hc.jsecs.org/irg/ws2021/app/index.php?id=02808>). The only written evidence for unification presented was based on data from the MOE *Dictionary of Variants* and SAT database in comment #1285. It seems to imply there are 7 unencoded characters listed but in fact 5 of the 7 are already encoded, albeit with 賁 as the right hand component. MOE C02303-002 𠂔女束貝 is encoded as U+218B2 𠂔, SAT 𠂔束貝 as U+23FD9 𠂔, MOE A02944-006 𠂔禾束貝 as U+258BC 𠂔 and MOE A03166-006 𠂔糸束貝 ~ 𠂔 A03166 as U+31E8A 𠂔 糸賁.

For 責 ~ 𠂔束貝:	
𠂔 B06006-003 ~ 𠂔 B06006	
𠂔女束貝 C02303-002 ~ 𠂔 C02303	
𠂔束貝 (SAT): 𠂔 ~ 𠂔, 𠂔 ~ 𠂔	
𠂔石束貝 (SAT): 𠂔 ~ 𠂔	
𠂔禾束貝 A02944-006 ~ 𠂔 A02944	
𠂔糸束貝 A03166-006 ~ 𠂔 A03166	
𠂔束貝 (SAT): 𠂔 ~ 𠂔	

(Figure 2: ws2021 comment #1285)

There are there are no UCS unification examples and, not surprisingly, further UCS disunification examples exist, such as:

U+22159 幘 巾賁	vs	U+5E58 幘 巾賁
U+228BB 幘 巾賁	vs	U+397D 幘 巾賁
U+23707 幘 木賁	vs	U+6A0D 幘 木賁
U+23A6C 殯 歹賁	vs	U+3C74 殯 歹賁
U+27894 覲 賁見	vs	U+468D 覲 賁見
U+288DC 醕 酉賁	vs	U+288A6 醕 酉賁
U+29F1C 鱣 魚賁	vs	U+9C3F 鱣 魚賁

The question of whether 賁 and 束貝 can be unified is not based on UCV #190d alone but rather the much older ws2017 level 2 UCV #422 of 束 and 束. Based on UCS precedent, 束貝 is unifiable with 賁 but not unifiable with 賁. Hence 賁 should be removed from UCV #190d.

The above shows the importance of considering the number of existing UCS unification and disunification examples involved before unifying characters or adding a UCV. It should be noted that non-UCS sources such as the MOE Dictionary of Variants or the SAT database whilst authoritative in some respects they use different criteria to UCS and so as such do not provide a reliable basis for deciding unification issues.

Other comments on the proposed rule have been made such as those on <https://hc.jsecs.org/irg/ws2021/app/index.php?id=01878>.

If, as suggested above, the proposed PnP addition is not suitable and when considering new unification issues there is need for extended discussion both written and in meetings, the question of IRG procedure is how to better manage them. An analysis of the data supplied in submissions to the current working set and a consideration of process to date suggests one possible addition to PnP that should smooth the progress of the next working set in that it would limit the proportion of characters submitted that are likely to require extended discussion, and this will be looked at in part 2.

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 , , , , , , , or and followed by two IRG approved character description components.

or

Short and unambiguous IDS: An IDS consisting of one IDC, either , , , , , , , or and followed by two IRG approved character description components, or 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 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%

Universal Multiple-Octet Coded Character Set
UCS

ISO/IEC JTC1/SC2/WG2/IRG N2612Feedback

Date: 2023-10-12

Source:	China
Author:	TAO Yang, CHEN Zhixiang
Title:	Feedback on IRGN2612
Meeting:	IRG #61
Status:	Member's submission
Actions required:	To be considered by IRG
Distribution:	IRG
Medium:	Electronic
Page:	2
Appendix:	Null

The motive of this discussion is meaningful.

1. The most ideal state for the transliteration of ancient Chinese characters is, when the academic community assigns a recognized classification font to each ancient Chinese character.

2. The Liding result of ancient script, stone carving script, and miscellaneous script did result in some accidental, redundant, and overly strict binding glyphs, and not every Liding glyph needs to be encoded.

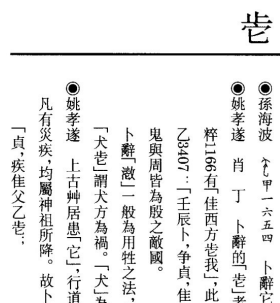
But this pending rule ignores the complexity of Chinese character culture and the difficulty of professional operation.

1. Liding is an uncertain behavior, and it is inevitable for ancient philologists of the same era to have differences in the method of Liding for the same ancient character. Various results may not necessarily compare a single correct option, and IRG cannot define the rules of transcription for the delineation of ancient characters in contemporary ancient philology based on their habitual understanding of commonly used regular script Chinese characters.

2. It is a common phenomenon in history to assign multiple forms to an ancient script, and it is also a faithful reflection of the process of script transformation and evolution. It does not have an idealized linear path, but rather in the complex evolution, multiple scripts influence each other, and ancient and modern literature influence each other. We should not use the current standardization intention to establish particularly precise concepts of right and wrong in the work of ancient people.

3. Even if an ancient script has multiple Liding script shapes and one of them is determined to be the most suitable Liding script scheme, other academic influential Liding methods still have preservation significance and should still be encoded for the convenience of academic use.

For example, 𪚩止它 is the type of Liding script that has already been used for decades in the field of ancient philology.



Professor Qiu Xigui pointed out that the lower part is 虫, which is why there is a new Liding script 𪚩止虫.

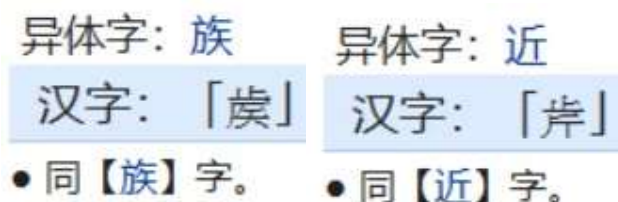
2227 𪚩	1. 𪚩 01663 (A4)	𪚩 02925 (A4)	𪚩 00095 (A7)	𪚩 00235 正 (A7)
	𪚩 00371 正 (A7)	𪚩 00371 正 (A7)	𪚩 00371 正 (A7)	
	𪚩 00440 正 (A7)	𪚩 00454 正 (A7)	𪚩 00454 正 (A7)	

So it is unimaginable that previously widely circulated characters appeared in various literature without being included.

4. If the Liding script has already been passed down in the literature, it has a certain dissemination significance, forming the value of textual research, and also generating supporting significance for the version system of the literature. This type of Liding script should also be preserved, rather than being encoded based on the correctness of the official script.

5. The right and wrong of the Liding script should still be judged by the philologist, and the coding work should be faithful to responding to the needs of various fields for the use of characters, rather than replacing the philologist in deciding which to retain and which to remove, and never acting on behalf of others.

6. If we hope that all ancient Chinese characters have a unique Liding script, then we not only overlook the historical differences between oracle inscriptions, bronze inscriptions, and small seal scripts, but also will remove some of the encoded Liding characters from the Kaishu script font.



Of course, the complexity involved goes far beyond that.

Title: Response to Feedback to IRGN2612
Source: Henry Chan
Date: 2023-10-17
Status: Individual Contribution to IRG #61
Action: For consideration by IRG
Pages: 3

Response to John Knightley's comments:

John suggests that 責 should be removed from UCV #190d, claiming that there are many existing disunification examples in Extension B. However, it has been agreed in IRG meetings that Extension B disunification examples are not considered as existing unification examples.

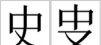

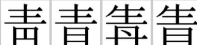
The quoting of the sources from MOE Dictionary is merely to prove that they are variants, and to prove the variation is systematic. It is not used as a direct basis for unification.

In recent IRG meetings, variants arising from using a different component form for the same radical have been changed from NUCV to UCV, since these differences are more appropriately unified and coded using IVS. For example NUCV #404, #405 has been changed to UCV, and new UCV #417, #430 has been added:

404 · unifiable 艸++  ++	405 · unifiable 攴攴  攴	417 · unifiable 𠂔𠂔 月𠂔 肉𠂔  𠂔	430 · unifiable 彡水  水
--	--	--	--

In essence these are calligraphic simplifications which are greatly generalizable and happened as part of the transformation between the Seal script, Clerical script, and Regular script.

These unifications are not limited to different forms of radicals. Besides 責, we also have other UCV examples where the “complex” or “full” form which corresponds more closely to the form in Seal script is now unified with the “common” form which is predominantly used in Regular script:

469 · unifiable 史史  史	202 · unifiable 敖敖  敖	319 · unifiable 青青青  青
---	---	--

My suggestion is that we codify this general class of unification so we can reduce the number of variants we need to code.

This is indeed a departure from the encoding model in Extension B, but also consider that the initial encoding model in URO does not include any of these “complex” or “full” forms, and none of the original source standards in the URO contain these forms.

Note, the “complex” form which is found in dictionaries is often not even consistent. Consider the following encoded forms:

勝 - 勝 勝 - 𠂔 勝 - 勝

vs

朕 - 𨾏

vs

膳 - 𨾏 膳 - 𨾏

In the first set, the 月 has been expanded to the original form 舟, while in the second set, the 关 has also been expanded to 𨾏, while in the third set, even the 𨾏 has been expanded to the full form 𨾏火収.

By encoding these variant forms as separate characters, it means users either need to already know the exact form used by a particular text, or the digitization system needs to maintain huge mapping tables for the variant characters back to the common characters. Both of which are unwieldy for use.

The fact is the general public do not even recognize these characters, so most digitization systems choose to convert the characters to the common ones, which results in an irreversible loss of information.

This is no longer an issue if the characters are unified and the variant forms are encoded as variants in an IVD collection. Any Unicode compliant text processing software should automatically handle variants encoded this way correctly, without the need for additional mapping tables and custom logic.

Response to Tao Yang and Chen Zhixiang's comments:

The goal of updating the IRG PnP to address unification regarding differences in transliteration is mainly to target the issue around one or more components taking a different form due to calligraphic simplification or component merging.

Similar examples are like 更 and 𠂇, 曹 and 𡩊, 晉 and 𡩊 etc. Note these “complex” or “full” forms have already been encoded as separate characters, so any new UCV only applies to characters which include them as components.

The goal of suggesting the update was not to block the encoding or suggest unifying characters like 𠂇止它 and 𠂇止虫, or 𠂇止矢 to be unified to 族.

Obviously the two types of transliteration are not the same.

The first one is the transliteration of one or more components into a single joined component, which arises due to the stroke simplification that happened between the development of the Clerical script and Regular script.

The second one is the transliteration of ancient characters where the character composition is not the same.

The feedback from Tao Yang and Chen Zhixiang also mentions “right or wrong should be judged by philologists”.

However, IRG should not be involved in the academic viewpoint of “right or wrong”. As long as a character has a required use, whether it is right or wrong, as long as it is not a printing error in limited distribution, it should be coded – either as a new character, or unified with an existing character.

If a character is unified it does not mean it cannot be coded. It only means it will not be coded as a new character under IRG, and it automatically means it can be coded as a variant in an IVD collection.

If philologists deem a particular “complex” or “full” form to be particularly noteworthy and correct according to some orthographic standard or academic theory, it does not mean that it needs to be coded as a separate character. Unification only needs to consider that two forms are interchangeable semantically, and submitters are only required to provide one representative glyph.

The matter of which form is “correct” or which form is “more correct” should be defined by the user community of these variants. If a given user community considers some forms are “more correct” than others, it can be represented by placing them into different IVD collections. This should be kept outside the scope of IRG.