ISO/IEC JTC1/SC2/WG2/**IRG N2791**

Date: 2025-03-03

Title: **Feedback to IRG N2734 on IRG's work**

Source: **Wang Xieyang (王谢杨), Center for Toponym Research of SISU (四川外国语大学地名研究中心)**

Status: **Individual Contribution**

Action: **For consideration by IRG**

Note: The Chinese translation of this document starts at page 5.

We basically agree with the points in the document IRG N2734 *Proposal on Revising PnP to Enhance the Work Efficiency of IRG*. Regarding related issues, we also have some overall ideas and specific suggestions.

1. IRG should cancel the submission quota limits for all member bodies

China is the only country in the world that uses Chinese as its official language, and it possesses the richest Chinese character documents globally. In recent years, with the rapid development of China's economy and culture, the demand for digitalizing documents written in Chinese characters has significantly increased. During the process of digitizing the documents, the number of unencoded characters discovered has also increased sharply, and many of these unencoded characters have extremely important uses. Currently, China's demand for encoding Chinese characters is not only the largest in quantity but also extremely urgent. The current quota of 1,000 characters per submission is far from meeting China's demand. For example, our center alone has currently collected more than ten thousand unencoded characters. If submitted according to the current quota, it will take 10 submission periods, approximately 20 to 30 years, which is completely unacceptable. Considering that the review tools are already relatively advanced and easy to use, we suggest canceling the submission quota limits for every member body and increasing the character number limit for future IRG working sets.

Specifically, we propose to modify the following content in IRG PnP(V17):

2.1.1.d.(1)

Original Text:

……the size of a collection or a part of an IRG collection, to be reviewed by IRG as a working set normally does not exceed 4,000 ideographs. Based on this principle, submitters should refrain from submitting more than 1,000 characters in each call for an IRG collection.

Suggested Text:

……the size of a collection or a part of an IRG collection, to be reviewed by IRG as a working set normally does not exceed 20,000 ideographs. ~~Based on this principle, submitters should refrain from submitting more than 1,000 characters in each call for an IRG collection.~~

3.1.e

Original Text:

……Each submitter is allowed to submit no more than 1,000 characters. As the normal work set size is set at 4,000, IRG will use the guidelines given in Annex L to estimate the number of working sets for the collection in case the total number of characters is much larger than 4,000……

Suggested Text:

……Each submitter is allowed to submit no more than 1,000 characters if the submitter did not request to change the quota before submission. As the normal work set size is set at 20,000, IRG will use the guidelines given in Annex L to estimate the number of working sets for the collection in case the total number of characters is much larger than 20,000……

Annex L: Guidelines for Forming Working Sets with an Upper Limit

Original Text:

……The current limit ($Limit_{IRG}$) is set to about 4,000 ideographs. Also, each submission should not go beyond 1,000 ideographs……

Suggested Text:

……The current limit ($Limit_{IRG}$) is set to about 20,000 ideographs. ~~Also, each submission should not go beyond 1,000 ideographs~~……

2. IRG should accelerate the review speed and focus on its own duties

The review speed of IRG is rather slow currently. Practical situations have shown that even the much-criticized CJKUI Extension B has a much less severe impact on applications than some experts have claimed. However, the harm caused by the lack of encoding and the lag in encoding is already quite significant. In China, many place names containing unencoded characters have been changed because of that, and the characters used in numerous classic works and even in the scientific field are also difficult to remain stable in information exchange. This has had an extremely profound negative impact on the development of Chinese culture, science, and technology. IRG has been discussing and trying to accelerate the review speed for a long time, but has not succeeded so far. If the review speed of IRG remains slow in the future, users will have to seek other solutions in order to safeguard their cultural rights and interests. We believe that, practically, IRG should no longer be overly concerned about certain characters, nor should it waste time on whether a certain character should be unified.

Firstly, the core responsibility of IRG is to encode Chinese characters, thereby providing the possibility of stably exchanging these characters between different devices. IRG or its experts have no right to determine whether a Chinese character is "correct" or "standard" and decide whether to encode a certain character based on this. Therefore, whether a character in a historical document is an error form should not be a matter for IRG to consider. If an error form is studied by someone or there is a need for information exchange, it has the value of being encoded. Furthermore, if a certain error form is submitted by a certain member body, it has the necessity to be encoded.

Secondly, IRG is not an academic research organization and there is no need to conduct detailed research on the pronunciation, meaning, and rationale of a character that is impossible to be unified. In order to fulfill its responsibilities, IRG only needs to confirm that the submitted character forms are consistent with the character forms in the evidence and simply verify the pronunciations, meanings, and rationales of the characters that may be unified.

Thirdly, considering the difficulty and extent of support for IVS and encoded characters by current devices and applications, the situation of characters is clearly superior to that of IVS, and this situation is unlikely to change for a long time in the future. Therefore, we believe that for variants that have the need to maintain differences in character forms within a relatively large scope (such as variants that are submitted relatively late but are not rare, and variants with important uses), they should be encoded separately, so as to safeguard the legitimate cultural rights and interests of countries within the Chinese character cultural circle.

Finally, in its practical work, IRG often starts solely from the perspective of philology experts and decides whether to unify two characters only based on philological theories. It frequently classifies y-variants in the eyes of most ordinary people as z-variants. However, ISO/IEC 10646 is actually a practical standard, and the majority of its users are ordinary people. If the decision on whether to unify Chinese characters is made only based on academic theories, it will cause trouble for most users. Therefore, we believe that IRG should take into account practical and non-theoretical factors and assign two different code points to potentially unifiable Chinese characters that have important uses and significant differences in character forms.

In summary, we propose the following:

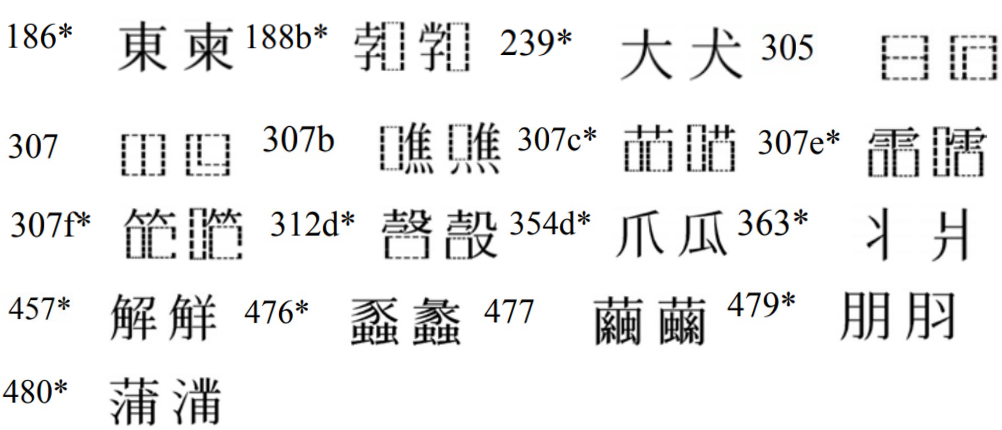
a. If a Chinese character has significant differences in its character form from the potentially unifiable characters and has important uses, and the submitting member body does not agree to unify this character, then this character should be encoded separately.

b. Even if a Chinese character can be unified according to Level 2 UCV, if the submitting member body wishes for it to be encoded separately, then this character should be encoded separately. If necessary, some current Level 2 UCV that are similar in character form and of the same origin can be classified as Level 1 UCV. For example, among the UCV168 (Version: Mar/30/2024), the first three variants and the last two variants can be regarded as each other's Level 1 UCV.

168*


c. Permit and support the proposals submitted by member bodies for disunifying the Chinese characters that were previously unified, based on the above-mentioned items a and b.

d. It is recommended to change the following UCV into NUCV(Version: Mar/30/2024).



3. IRG should earnestly respect the cultural sovereignty of each country

We believe that determining the character usage and related norms of one's own country is an important part of national cultural sovereignty. IRG should fully respect the cultural sovereignty of each member state and earnestly attach importance to the character usage requirements submitted by governments of member states. It should not, on the pretext of "saving code points" or "avoiding errors," force governments of member states to provide relevant information such as pronunciations and meanings of characters used in key fields like administrative affairs, science and technology, and culture. Taking characters used in the administrative field as an example, refusing to encode the characters submitted by governments of member states on the grounds of the lack of pronunciations and meanings is undoubtedly a disregard for and an infringement upon the national cultural sovereignty. IRG must avoid the recurrence of such situations.

From a non-political perspective, once a character is incorporated into the government's public administrative system, it will come into contact with the general public. Subsequently, it will appear in various documents and archives, and thus, there will be an actual need for information exchange using this character. Therefore, even if this character is an error form, based on the actual needs of information exchange, it should still be encoded.

Based on the viewpoints above, we suggest that IRG recognize the characters used in government affairs, science and technology, and cultural classics submitted by member states as authoritative evidence. Moreover, as long as there are no potentially unifiable characters encoded, governments of member states do not need to provide pronunciations and meanings for the characters used in government affairs. If our suggestions can be adopted, the relevant content in Section 2.2.1.d.(2) of the IRG PnP should also be adjusted accordingly.

(This is the end of the proposal in English. Following is the Chinese translation.)

我们基本同意 IRG N2734（《关于修订原则与程序以提高表意文字工作组工作效率的提案》）件中的提议，对于相关问题，我们还有一些整体性的想法和具体的建议。

1. IRG 应取消各提交方的提交额度限制

中国是全球唯一以汉语作为官方语言的国家，拥有全世界最为丰富的汉字文献资源。近年来，随着中国经济与文化的飞速发展，对电子化汉字文献的需求大幅增加。在电子化汉字文献的过程中，发现的未编码字数量也急剧增多，其中许多未编码字都具有极其重要的用途。目前，中国对编码汉字的需求不仅数量庞大，而且极为迫切，当前 1000 个/提交期的额度远远无法满足中国的用字需求。例如，仅我们中心目前已收集到的一万余个未编码字，若按此额度提交，就需要 10 个提交期，大约耗时 20 至 30 年，这显然是不可接受的。考虑到当前的审核工具已经较为先进且易用，我们建议取消各提交方的提交额度限制，并提高之后 IRG 工作集的字数额度。

具体而言，我们建议修改《表意文字工作组原则和程序（第 17 版）》中的以下内容：

2.1.1.d.(1)

原文：

……the size of a collection or a part of an IRG collection, to be reviewed by IRG as a working set normally does not exceed 4,000 ideographs. Based on this principle, submitters should refrain from submitting more than 1,000 characters in each call for an IRG collection.

修改后：

……the size of a collection or a part of an IRG collection, to be reviewed by IRG as a working set normally does not exceed 20,000 ideographs. ~~Based on this principle, submitters should refrain from submitting more than 1,000 characters in each call for an IRG collection.~~

3.1.e：

原文：

……Each submitter is allowed to submit no more than 1,000 characters. As the normal work set size is set at 4,000, IRG will use the guidelines given in Annex L to estimate the number of working sets for the collection in case the total number of characters is much larger than 4,000……

修改后：

……Each submitter is allowed to submit no more than 1,000 characters if the submitter did not request to change the quota before submission. As the normal work set size is set at 20,000, IRG will use the guidelines given in Annex L to estimate the number of working sets for the collection in case the total number of characters is much larger than 20,000……

Annex L: Guidelines for Forming Working Sets with an Upper Limit

原文：

……The current limit ($Limit_{IRG}$) is set to about 4,000 ideographs. Also, each submission should not go beyond 1,000 ideographs……

修改后：

……The current limit ($Limit_{IRG}$) is set to about 20,000 ideographs. ~~Also, each submission should not go beyond 1,000 ideographs~~……

2. IRG 应加快审核速度并专注于本职工作

IRG 目前的审核速度较为缓慢。实际情况表明，即使是饱受争议的中日韩表意文字扩展 B 区，其对实际应用产生的影响也远没有部分专家所宣称的那么严重。然而，汉字无编码以及编码滞后所带来的危害却已十分明显。在中国，许多包含生僻字的地名因此被迫更改，众多经典典籍中的用字乃至科研领域的用字也难以在信息交换中保持稳定，这对中国文化和科技的发展都产生了极为深远的负面影响。IRG 长期以来一直在讨论并尝试加快审核速度，但至今仍未能成功。倘若未来 IRG 的审核速度依旧缓慢，用户将不得不寻求其他解决途径以维护自身的文化权益。我们认为，在实际工作中，IRG 不应再过度纠结于单个汉字的取舍，也不应在某个单字是否应被统合的问题上浪费时间。

首先，IRG 的核心职责是对汉字进行编码，以实现这些字在不同设备间的稳定交换。IRG 及其专家无权判定一个汉字是否“正确”或“规范”，并据此决定是否对该字进行编码。因此，一个历史文献中的字是否为讹字不应成为 IRG 考虑的因素，只要讹字有人研究，且存在信息交换的需求，就具有编码的价值。进一步来讲，若一个讹字被某个提交源提交，它就具备了编码的必要性。

其次，IRG 并非学术研究组织，无需对一个不可能被统合的字的音、义、字理等进行细致深入的研究。为了履行其职责，IRG 只需确定提交的字形与证据中的字形一致，并简单核实可能被统合的字的音、义、字理等即可。

再次，从当前设备和应用对 IVS 及字符的支持难度及广泛程度来看，字符的情况明显优于 IVS，并且在未来很长一段时间内，这种情况都难以改变。因此，我们认为，对于那些在较大范围内有保持字形差异需求的异体字（如提交较晚但并非罕用的异体字、有重要用途的异体字），应将它们分别编码，以保障汉字文化圈各国的正当文化权益。

最后，IRG 在实际工作中，往往单纯地从文字学专家的角度出发，仅依据文字学理论来决定两个汉字统合与否，常常将大多数普通人眼中的 y 变体归类为 z 变体。但 ISO/IEC 10646 实际上是一个实用性标准，其用户大多数还是普通人。如果仅依据学术理论来决定汉字统合与否，就会给大多数用户带来困扰。所以，我们认为，IRG 应将现实性的、非理论性的因素纳入考量范围，给有重要用途且字形差异较大的两可统合字两个不同的编码。

综上，我们建议：

1、若某汉字与潜在的可统合字字形差异较大且有重要用途，且提交方不同意该字被统合，则该字应单独编码。

2、即使某汉字可依据第二级可统合部件变体（UCV）被统合，若提交方希望单独编码，则该字应单独编码。如有必要，可以将现行第二级可统合部件变体中部分字形相近且同源的变体划分为第一级可统合部件变体。如，可统合部件变体 168 中（Version: Mar/30/2024），前三个变体和后两个变体可互为第一级可统合部件变体：

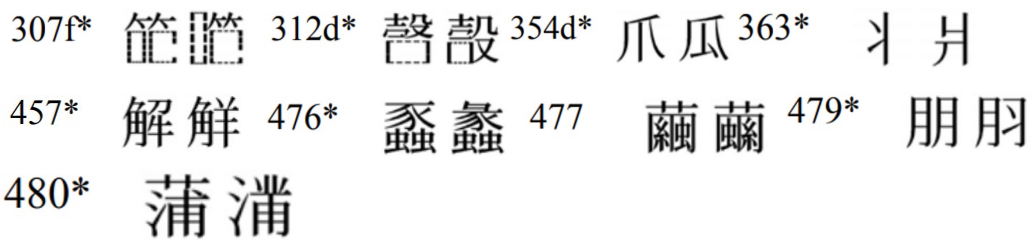
168* 臼 臼 臼 凹 血

3、允许并支持各提交方基于上述第 1、2 条提出单独编码之前被统合的汉字。

4、建议将以下可统合部件变体改为不可统合部件变体（NUCV）。

186* 東 東 188b* 剪 剪 239* 大 犬 305 日 日

307 口 口 307b 嗤 嗤 307c* 苗 苗 307e* 靐 靐



3. IRG 应切实尊重各国的文化主权

我们认为，确定本国的用字及相关的规范属于国家文化主权的重要范畴。IRG 应当充分尊重各成员国的文化主权，切实重视各国政府所提交的用字需求，而不应以“节省码位”或“避免错误”等为由，强行要求各国政府提供政务、科技以及文化等关键领域用字的读音、含义等相关信息。以政务用字为例，以未提供音义信息为由拒绝编码各国政府所提交的政务用字无疑是对国家文化主权的漠视与侵犯，IRG 务必避免此类情况再次出现。

从非政治层面考量，一个字一旦被纳入政府的公共政务系统，便会与广大民众产生接触，进而会出现在各类文件、档案之中，随之也就产生了运用该字进行信息交换的实际需求。因此，即便这个字属于讹字，基于信息交换的实际需要，也应当编码该字。

基于上述观点，我们建议 IRG 将各国政府提交的政务、科技以及文化经典等领域的用字认定为权威证据。并且，只要不存在可以统合的已编码异体字，各国政府就无需为政务用字提供读音和含义等信息。倘若我们的建议能够获得采纳，《表意文字工作组原则和程序》2.2.1.d.(2)节的相关内容也应作出相应调整。

(End of Doc)