**ISO/IEC JTC1/SC2/WG2 IRG N2866R**

**DATE: 2025-10-1**

**Title: Script-Hybrid Characters and GB 18030**

**Source: Witty Wen (文辰奘)**

**Status: Individual Contribution to IRG #65**

**Action: For Consideration by China, UTC and IRG**

## I. Introduction

In recent IRG meetings, a set of “script-hybrid” characters – characters that combine Han components with letters or kana – has sparked debate. Examples include ⿰广 K and ⿰广 O (abbreviations of the Japanese name Keiō as 慶應), as well as a character in the upcoming CJK Extension J defined as ⿰X 也. These hybrids incorporate non-Han letters (Latin “K”, “O”, “X”, etc.) as structural elements intended to convey pronunciation. IRG previously concluded that such items do not belong in the CJK Unified Ideographs in #61. The topic was revisited at IRG #63 and #64 and consumed significant agenda time. In Unicode 17.0, the case now encoded as U+323BF (WS2021 UK-20538) was accepted; in the final code chart its left component is rendered as the Han component 乂, not a Latin ‘X’. However, it should trace to a Latin 'X', from an intent standpoint. These developments have prompted active discussion but no consensus on a stable way forward. In the meantime, these forms—especially those that combine Latin or hiragana with Han components—are categorically distinct from CJK Unified Ideographs; accordingly, in this document I treat them as characters rather than ideographs.

I share the concerns raised by China and others regarding script-hybrid characters. In this proposal, I propose a practical solution: script-hybrid characters should be placed in a separate block rather than within the unified CJK ideographs. As an alternative, I also suggest a fallback option for China to preserve consistency if such characters are nevertheless added to CJKUI. The aim is to reach consensus at IRG #65 and avoid further prolonged debate.

## II. Script-Hybrid Characters and CJK Unified Ideographs

Characters that combine Han components with Latin, kana, or other scripts challenge the fundamental definition of what counts as a "Hanzi." China has repeatedly emphasized this point: while such hybrids may function as ideographs in practice, they exceed the established understanding and technical definition of Han characters. Including them indiscriminately in the CJK Unified Ideographs (CJKUI) would blur the line between alphabetic scripts and Han, undermining both sinological theory and practical assumptions in computing.

China's persistence in raising this concern within IRG reflects a legitimate and necessary defense of ISO/IEC 10646's scope, and this effort deserves recognition. At the same time, the repeated debates have consumed substantial meeting time and created difficulties for IRG's overall progress.

A concrete case illustrates the boundary: the character now encoded as U+323BF (UK-20538). While some sources wrote it as 'X 也', the encoded reference glyph uses the Han component 乂 on the left. Some fonts may stylize it with Han-like strokes, but semantically it remains the Latin letter. This is categorically different from characters such as 刈 or 艾, whose shapes only coincidentally resemble Latin letters but are in fact established Han radicals. This gap between source intent and encoded realization highlights the need for clear principle. The □X 也 case also serves as a reminder that China needs to consider carefully how such characters should be treated, particularly in relation to GB 18030.

**Side Note:** I am cautious about embedding gender politics in encoding process. The "X 也" case is not a commonly attested character and lacks historical pedigree; in hindsight, the review may not have fully weighed these sensitivities. It has evidently touched on gender-related debates and attracted criticism. As a matter of prudence, we should avoid bringing such issues into the encoding process wherever possible.

The current IRG Principles and Procedures (P&P) were designed exclusively for Han ideographs. They contain no provisions for treating non-Han letters as components, for stroke-

counting alphabetic shapes, or for extending IDS syntax. Attempting to adapt P&P in this way would create complexity and inconsistency, blur the conceptual boundary between Hanzi and alphabetic scripts, and impose unnecessary burdens on standards bodies and vendors.

For these reasons, while stakeholders' positions must be respected, the issue requires a clear resolution. The following section presents one practical solution — establishing an independent 'CJK Hybrid Characters' block to address the scope expansion.

## III. Independent Block for Script-Hybrid Characters

I believe the encoding of script-hybrid characters is both reasonable and meaningful. A feasible solution is to encode them in a dedicated block specifically for "Script-Hybrid Characters" (i.e. named 'CJK Hybrid Characters'). UTC has suggested the possibility of defining such a block on Plane 1 (the SMP), and China has also agreed in principle to this approach. This would segregate characters like ⿸广 K, ⿸广 O, and other Han–Latin, Han–Kana, or Han–Hangul combinations into their own category, distinct from true CJK unified ideographs.

Establishing a separate block offers several advantages:

- Maintaining Clarity of Definition: It ensures that any character in the CJK Unified Ideographs blocks (URO and Extensions A–J/future) is composed exclusively of Han strokes or radicals. Anything involving Latin letters, Japanese kana, Korean hangul, Zhuyin (Bopomofo), etc., would reside outside that range. This avoids confusion for users and implementers about what constitutes a Han character.
- Preserving IRG Processes: By isolating these cases, IRG can update its procedures in a focused way for the new block, without overhauling the core P&P for unified ideographs. Issues such as stroke counting for letters or extending IDS syntax for non-Han components could be addressed within this block. The block's naming could explicitly indicate its connection to CJK while marking its special status (e.g., "CJK-Hybrid-Characters").
- Cohesive Treatment: All script-hybrid characters used in the broader CJK writing sphere could be encoded in one block. This includes not only Han–Latin forms but also Han–

Kana, Han–Hangul, and others. By handling them together, we ensure consistent criteria for inclusion and avoid one-off exceptions in the main CJK set.

- Implementation & Governance. Submission and technical review remain under IRG procedures (evidence vetting, unification, glyph review). UTC coordinates hosting and publication of the independent 'CJK Hybrid Characters' block. For data interoperability, include these entries in Unihan (e.g., kTotalStrokes, kRSUnicode). Rather than overloading the Unicode Script property, introduce a metadata tag 'CJK-Hybrid' to flag such entries while keeping them discoverable alongside CJKUI.

The recent push to accept these hybrids as "ordinary" ideographs is understandable in terms of simplifying Unicode processing. However, there remains a general lack of study, and no consensus exists on their structure, radicals, stroke counts, variants, or unification rules. Rushing to encode them as regular ideographs is too risky. By moving them to a separate block, we gain time to develop proper guidelines without disrupting the ongoing work of CJK extensions. This solution is worth serious consideration.

**Side Note** (Gray Area): For katakana-shaped, bopomofo-like, or hangul-like components that may be manifested as Han strokes: if IRG cannot reach consensus, default future proposals to the independent block; China may decide mapping for already-encoded legacy cases. For new proposals, prefer the independent block by default, while allowing case-by-case discussion.

## IV. Additional Issue for China's Consideration

The relationship between Unicode/ISO 10646 and China's national standard GB 18030 has historically been one of close alignment. GB 18030 includes the entire repertoire of Unicode CJK Unified Ideographs, ensuring that any Han character encoded in UCS is supported in Chinese systems. This cooperation has worked well for several decades.

However, if the proposal to establish a separate block is not accepted by IRG, then from China's perspective such characters clearly "exceed the scope of technical processing of Hanzi," and it

would be reasonable for China not to accept them into its national standard. I strongly recommend that China, based on its own practical needs, address the issue by updating GB 18030 rather than continuing endless debates at the IRG or higher level.

Specifically, if script-hybrid characters are encoded (contrary to the advice above) as part of CJK Unified Ideographs, I recommend that China consider leaving those code points unassigned (holes) in the GB 18030 repertoire. In practice, this would mean that characters such as ⿰X也 (already in Extension J), and any future hybrids such as ⿰广K or ⿰广O, would not be mapped in GB 18030 even though they have Unicode code points. This "hole punching" approach would signal that these are not recognized as standard Chinese characters domestically, and it would avoid complicating Chinese-language implementations with characters that do not meet the Han definition.

That said, the effect of such a move would be significant. It would break the one-to-one correspondence between Unicode CJK and GB 18030 for the first time, something all parties have so far preferred to avoid. I consider it important to raise this prospect: if the majority within IRG chooses to encode script-hybrids characters directly into CJK extensions, they should recognize that this could erode the universal acceptance of the UCS repertoire in China. No one wants a scenario where end-users encounter a "defined" character that is unsupported in Chinese environments due to standard misalignment. Thus, if necessary, treating these characters as unmapped in GB 18030 remains a valid fallback option.

## V. Conclusion and Requested Actions

The discussion of script-hybrid characters highlights both the practical needs of users and the importance of maintaining a clear scope for CJK Unified Ideographs. These characters do exist, but their inclusion within CJKUI raises unresolved questions of definition, procedure, and implementation.

- Creating a separate block is the most balanced and forward-looking solution. It allows these forms to be encoded without altering the Han-only scope of CJKUI and gives space for tailored procedures to be developed.
- Leaving hybrid code points unmapped in GB 18030 could serve as a pragmatic fallback if hybrids are nevertheless placed in CJK extensions. This would keep China's implementation consistent.

**Action Requested:**

- IRG: I recommend affirming the principle that CJK Unified Ideographs remain composed entirely of Han components and giving due consideration to establishing an independent block named 'CJK Hybrid Characters'.
- UTC: I recommend, in collaboration with IRG, exploring the technical feasibility of defining the independent 'CJK Hybrid Characters' block and developing minimal rules for it.
- China: I recommend considering, if script-hybrid characters are nevertheless added to CJKUI, updating GB 18030 to leave those code points unmapped as a fallback approach, so as to avoid further prolonged debates at the IRG level.
- Terminology: I recommend formalizing the convention that 'script-hybrid' denotes characters rather than ideographs, and maintaining this distinction consistently in IRG documents.

## Acknowledgement

Thanks to Dr. Ken Lunde and Mr. Tao Yang for reviewing this document and to Mr. Ao Weijun for a cross-audience perspective.