Title: Industry Recommendations for GB 18030 Testing & Certification
To: China Electronics Standardization Institute (CESI)
From: Unicode Technical Committee (UTC)
Date: 2023-01-17

The Unicode Technical Committee discussed testing and certification in relationship to the implementation of the GB 18030-2022 standard, which involved major industry developers of operating systems and fonts, and would therefore like to recommend changes to the testing and certification procedures that would benefit all customers.

For the purposes of this document, all references to the Unicode Standard equally apply to the ISO/IEC 10646 and GB/T 13000 standards.

Transcoding Recommendations

Among the 1,112,064 code points that transcode between the GB 18030 standard and the Unicode Standard, only 36 mappings were changed in the GB 18030-2022 standard. For both standards, there is a preferred code point for each of the affected 18 characters. Transcoding **between** the preferred code points of both standards is **bidirectional**. In order to migrate existing customer data to the preferred code points of both standards, we recommend that the transcoding of non-preferred points be **unidirectional** by transcoding **from** the non-preferred code points **to** the preferred code points. The following table provides an example, with preferred code points highlighted in **red**, and with arrows that indicate transcoding directionality:

Previous Mappings

GB 18030-2005		Unicode Standard
0xA6D9	⟷	U+E78D
0x84318236	⟷	U+FE10

Current Mappings

GB 18030-2022		Unicode Standard
0xA6D9	⟷	U+E78D
0x84318236	⟷	**U+FE10**

Recommended Mappings

GB 18030-2022		Unicode Standard
0xA6D9	⟵	U+E78D
0x84318236	⟶	**U+FE10**

The recommended mappings provide an optimal path for migration of existing customer data to preferred code points. In contrast, the current mappings provide no benefit for migration of existing customer data. For example, if an existing document is encoded according to the Unicode Standard and includes the character at code point U+E78D, whose use is now discouraged, transcoding to GB 18030 encoding using the current mapping will result in 0x84318236, which is not the preferred code point. But, if the recommended mapping were used, then U+E78D would be transcoded to the now-required—and preferred—code point, 0xA6D9. By implementing the recommended mappings, every transcoding from the Unicode Standard to GB 18030 encoding will result in fewer GB 18030 four-byte code points in customer data, and every transcoding from GB 18030 encoding to the Unicode Standard will result in fewer Unicode PUA (*Private Use Area*) code points in customer data.

The *gb18030-mapping-change-recommendations.txt* data file, which is a PDF attachment, provides the recommended mappings for all 36 code points.

Font Recommendations

With regard to the GB 18030 testing requirement to not display glyphs for 1) the 18 PUA code points that formerly mapped to preferred GB 18030 code points (see Table D.1 on pp 546 and 547 of the GB 18030-2022 standard); 2) the six PUA code points whose preferred GB 18030 code points changed (see Table D.2 on page 547 of the GB 18030-2022 standard); and 3) the nine CJK Compatibility Ideograph code points that were removed from the required portion of GB 18030 (see Table D.3 on page 548 of the GB 18030-2022 standard), please be aware that a long-standing industry practice among font developers is to never remove mappings from fonts. This is meant to ensure that existing documents display as originally intended. If the **Transcoding Recommendations** in this document were to be implemented, any requirement to not display glyphs for the affected PUA and CJK Compatibility Ideograph code points is neither practical nor necessary. Furthermore, application of any of the four normalization forms—NFC, NFD, NFKC, and NFKD—will result in the nine CJK Compatibility Ideographs becoming their canonical equivalents, which are CJK Unified Ideographs.

In lieu of requiring implementations to not display glyphs for PUA and CJK Compatibility Ideograph code points, a better approach would be to treat this as an IME (*Input Method Editor*) problem, not a font problem.

In other words, our recommendation is to lift the requirement to not display glyphs for the affected PUA and CJK Compatibility Ideograph code points, which is neither practical nor necessary, and instead to ensure that IMEs emit only the preferred code points for the affected characters.

That is all.