L2/22-274R

Disruptive Changes in GB 18030-2022

Peter Constable
October 28, 2022
Revised December 9, 2022

Summary

A new edition of the GB 18030 standard was published earlier this year by China's national standards body responsible for character encoding standards, Chinese Electronics Standardization Institute ("CESI"). The new standard (designated GB 18030-2022) introduces changes from the previous edition (GB 18030-2005) that are disruptive for implementations that conform to Unicode while also trying to conform to the CESI standards. This document provides an overview of these disruptive changes.

The new edition also introduces major changes that are not disruptive, bringing GB 18030 up to date with more recent editions of ISO/IEC 10646 and more recent versions of Unicode. For example, the new edition of GB18030 adds CJK ideograph characters from CJK Extension D to CJK Extension F. These changes in the new edition are not discussed further in this doc.

Suggested UTC action

This document is provided for UTC awareness and discussion, as well as for awareness among Consortium members. No specific UTC action is requested, though feedback to CESI might be considered.

Background

GB 18030 is a national standard with stringent conformance requirements that regulate eligibility for products or services to be sold in China. The new edition defines three different levels of conformance requirement ("implementation levels") that may need to be satisfied, depending on the nature of the product or service and the customer audience.

- Level 1 applies to all products with requirements that include support for the URO block (CJK Unified Ideographs) and for CJK Extension A.
- Level 2 applies to "system software and support software… include[ing] operating system, database management system and middleware". Level adds a requirement to support all characters in the "Standard Chinese Characters List" (detailed in Annex E); that list has 8105 characters, most of which are in the URO, but also includes some characters from CJK Extensions A through F.
- Level 3 applies to all products "for government affairs services and public services", and adds the requirement to support all CJK ideographs up through Extension F, as well as Kangxi radicals.

GB 18030 defines a distinct, multi-byte encoding form that is (roughly) an extension of GB 2312, and in which characters are encoded in code unit sequences of one, two or four bytes. So, for example, some characters will be cited as a single byte value, such as "41" (hex); others, as a double-byte value, such as "A6D9"; and others, as a four-byte value, such as "84318236".

The character set is also defined by reference to GB/T 13000, which is China's national re-publication of ISO/IEC 10646, which in turn is equivalent in character repertoire and encoding to Unicode.

Because characters are defined in terms of both the GB 18030 encoding form and also GB/T 13000, a mapping is thereby defined between the GB 18030 encoding and Unicode/10646 encoding. To some extent, the encoding mapping could be defined by numeric computation over byte ranges, but there are many exceptions requiring per-code-point lookup tables.

Three classes of disruptive change

Annex D of GB 18030-2022 describes three specific classes of change from GB 18030-2005 involving 33 different characters and 55 code positions.

- The 2022 edition makes 36 changes in the mapping relationship between GB 18030 and Unicode encodings.
- The 2005 edition included 6 characters with double mappings. The 2022 edition removes the double mappings.
- The 2005 edition included 9 characters from the CJK Compatibility Ideographs block. In Unicode/10646, these all have canonical decomposition mappings to characters in the URO. In the 2022 edition, these nine compatibility characters are removed.

The first class comprises changes that are the most potentially disruptive. **Note that these have potential impact for ICU** and other encoding-conversion implementations.

The second and third classes of change have potential impact for fonts, input methods and user content. Given the characters involved, the risks are not as significant as for the first class.

Changes to Unicode encoding mappings

Annex D.1 of GB18030-2022 details "18 adjusted GB/T 13000 code positions". These involve 18 characters, but in fact affect mappings for 36 code points in GB 18030 and Unicode.

The characters in question were not yet assigned in Unicode / 10646 when GB 18030-2005 was developed. In that national standard, they were assigned code positions in "double-byte area 1" (not private use), with mappings to Unicode PUA code points in the BMP.

All these characters were subsequently standardized in Unicode 4.1 with code points either in the Vertical Forms block (FE10..FE19) or the CJK Unified Ideographs block (9FB4..9FBB).

In GB 18030-2022, changes for these characters reflect that they now have standardized code points in Unicode/10646 and GB/T 13000. However, the code positions in GB 18030 encoding were not changed. Rather, it was the mapping from GB 18030 encoding to Unicode that was changed.

For example, consider the character "  ", assigned since Unicode 4.1 as U+9FB4: here is it's code position in GB 18030-2005, in GB 18030-2022, and the mappings in each of those to GB/T 13000 / Unicode / 10646:

Code position in GB 18030-2005 and in GB 18030-2022	GB 18030-2005 mapping to Unicode	GB 18030-2022 mapping to Unicode
FE59	U+E81E	U+9FB4

For each character, there is a reciprocal change affecting the mapping of a different GB 18030 code position:

Code position in GB 18030-2005 and in GB 18030-2022	GB 18030-2005 mapping to Unicode	GB 18030-2022 mapping to Unicode
82359037	U+9FB4	U+E81E

A complete list of these changes is provided in Annex A, below.

GB 18030-2022 could have maintained stability in the encoding mapping to Unicode, and simply removed the chart glyphs from certain code positions and moved the characters to new code positions. Note that there are other cases in the 2022 edition in which code positions are "un-assigned".

But instead of maintaining stability in the mapping to Unicode, GB 18030-2022 maintains consistency in the GB 18030 encoding of these characters, even though there is little indication that data is created or interchanged using the GB 18030 encoding.

These changes have various potential, disruptive impacts.

First, for encoding conversion mappings in ICU and other implementations, there is a "catch-22" problem: If implementations are updated to reflect changed mappings, that could be a breaking change for existing applications that use the GB 18030 encoding, or for existing content encoded using the 2005 specification. But if encoding conversion implementations are *not* changed, that also could break interoperability with new implementations that interchange data using the 2022 specification.

Font implementations will also be impacted, with corollary impacts for end users. Most or all fonts today use Unicode encoding. Starting in August 2023, new product certification tests will be enforced, and the tests will require that fonts do not have any glyphs assigned to the PUA code points that were used in the 2005 specification. It is not difficult for font developers to map glyphs to different code points, but that change will break existing content created from 2005 to 2022 using those PUA characters. **This includes content created using Unicode encoding** as well as GB 18030 encoding.

There are other disruptive impacts affecting users involving IMEs and search.

A potential mitigation of these risks might exist if the 18 characters in question are infrequently used in content or applications. That is a plausible situation, though it is not easy to determine with any confidence.

Removal of PUA assignments for double-mapped characters

In GB 18030-2005, 6 characters were assigned two different code positions in the GB 18030 encoding: one in a private-use area, plus another in a non-PUA four-byte area. For each pair, both code positions were given a mapping to Unicode: a PUA code point in the BMP, plus a code point in CJK Extension B.

In GB 18030-2022, the 6 private-use code positions are no longer assigned any glyph. The Unicode mapping for those private-use code positions is not changed; only the character assignment is removed.

A complete list of these changes is provided in Annex B, below.

This has potential disruptive impacts for fonts and existing content. As noted above, certification tests beginning August 2023 will require that fonts do not have any glyphs assigned to the PUA code points that were assigned a glyph in the 2005 specification. Removing glyph mappings in fonts will break existing content created from 2005 to 2022 using those PUA code points. **This includes content created using Unicode encoding** as well as GB 18030 encoding.

There are other disruptive impacts affecting users involving IMEs and search.

A potential mitigation of risks might exist if the 6 characters in question are infrequently used. Given the characters in question, this seems plausible. Also, the fact that these characters had been given standard code position assignments as well as PUA suggests reduced likelihood that the PUA characters have been used in practice. It is not easy to determine actual usage in content or applications with certainty, however.

Removal of CJK compatibility characters

In GB 18030-2005, 9 CJK compatibility ideograph characters were included along with the corresponding non-compatibility character. These were assigned to distinct GB 18030 code positions. A Unicode mapping for each was defined: the 9 compatibility characters mapped to code points in the CJK Compatibility Ideographs block (U+F900..U+FAFF). In each case, the corresponding non-compatibility character mapped to the URO character that is the canonical decomposition for the character in the CJK Compatibility Ideographs block.

In GB 18030-2022, the 9 code positions previously assigned compatibility ideographs have been unassigned. A complete list of these changes is provided in Annex C, below.

As described in the previous section, there is potential disruptive impact for fonts and existing content, and also for IMEs and search. Potential mitigations regarding existing content described in the previous section also apply.

There is a significant difference, however. As yet, it is unclear what expectations will be enforced for these compatibility characters in certification testing starting in August 2023:

> *Since the code positions in GB 18030-2022 no longer have a glyph assigned, will certification tests require fonts to not have glyphs assigned for those characters?*

If that is the expectation enforced, then that will be in **direct conflict with Unicode**, which would deem such fonts to be conformant (barring other considerations).

Annex A: Full list of Unicode mapping changes from Annex D.1

The following is the full list of 36 encoding mapping changes in GB 18030-2022 versus the 2005 edition. For each pair of rows, mappings to Unicode code points are swapped.

Code positions in GB 18030 (2005 and 2022 editions)	Chart glyph in GB 18030-2022	GB 18030 -2005 mapping to Unicode	GB 18030-2022 mapping to Unicode
A6D9	’	U+E78D	U+FE10
84318236	(None)	U+FE10	U+E78D
A6DA	°	U+E78E	U+FE12
84318238	(None)	U+FE12	U+E78E
A6DB	`	U+E78F	U+FE11
84318237	(None)	U+FE11	U+E78F
A6DC	：	U+E790	U+FE13
84318239	(None)	U+FE13	U+E790
A6DD	；	U+E791	U+FE14
84318330	(None)	U+FE14	U+E791
A6DE	！	U+E792	U+FE15
84318331	(None)	U+FE15	U+E792
A6DF	？	U+E793	U+FE16
84318332	(None)	U+FE16	U+E793
A6EC	﹀	U+E794	U+FE17
84318333	(None)	U+FE17	U+E794
A6ED	﹏	U+E795	U+FE18
84318334	(None)	U+FE18	U+E795
A6F3	：	U+E796	U+FE19
84318335	(None)	U+FE19	U+E796
FE59	𠂇	U+E81E	U+9FB4
82359037	(None)	U+9FB4	U+E81E
FE61	𠂢	U+E826	U+9FB5
82359038	(None)	U+9FB5	U+E826
FE66	𠂤	U+E82B	U+9FB6
82359039	(None)	U+9FB6	U+E82B
FE67	𠂋	U+E82C	U+9FB7
82359130	(None)	U+9FB7	U+E82C
FE6D	𠂉	U+E832	U+9FB8
82359131	(None)	U+9FB8	U+E832
FE7E	𠂆	U+E843	U+9FB9
82359132	(None)	U+9FB9	U+E843
FE90	卓	U+E854	U+9FBA
82359133	(None)	U+9FBA	U+E854
FEA0	纟	U+E864	U+9FBB
82359134	(None)	U+9FBB	U+E864

Annex B: Removed PUA assignments

The following are complete details regarding 6 characters that were given an alternate PUA assignment in GB 18030-2005 but for which the PUA assignments have been removed in GB 18030-2022.

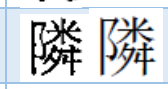
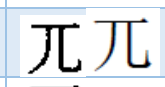
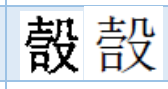
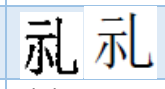
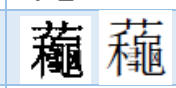
Unicode code point	Code positions in GB 18030 (2005 and 2022 editions)	Chart glyph in GB 18030-2005	Chart glyph in GB 18030-2022
U+E816	FE51	𠂇	(none)
U+20087	95329031	𠂇	𠂇
U+E817	FE52	𠂉	(none)
U+20089	95329033	𠂉	𠂉
U+E818	FE53	𠃌	(none)
U+200CC	95329730	𠃌	𠃌
U+E831	FE6C	𡈼	(none)
U+215D7	9536B937	𡈼	𡈼
U+E83B	FE76	𢦏	(none)
U+2298F	9630BA35	𢦏	𢦏
U+E855	FE91	𣲘	(none)
U+241FE	9635B63O	𣲘	𣲘

Annex C: Removed CJK compatibility ideographs

The following are details regarding 9 CJK compatibility ideographs removed in GB 18030-2022. Each pair of consecutive rows shows a compatibility ideograph and its corresponding canonical-decomposition ideograph. Note that, in some cases, there is a noticeable difference in glyphs in the GB18030 charts.

Unicode code point	Unicode canonical decomposition	Code positions in GB 18030 (2005 and 2022 editions)	Chart glyph in GB 18030-2005[1]	Chart glyph in GB 18030-2022
U+F92C	≡ U+90CE	FD9C	郎郎	(none)
U+90CE	n/a	C0C9	郎	郎
U+F979	≡ U+51C9	FD9D	凉凉	(none)
U+51C9	n/a	C1B9	凉	凉
U+F995	≡ U+79CA	FD9E	季季	(none)
U+79CA	n/a	B66A	季	季
U+F9E7	≡ U+88CF	FD9F	裏裏	(none)

[1] Rows for compatibility ideographs show two glyphs: the first is taken directly from GB 18030-2005, while the second is the glyph "in GB 18030-2005" as reported in GB 18030-2022.

U+88CF	n/a	D159		
U+F9F1	≡ U+96A3	FDA0		(none)
U+96A3	n/a	EB4F		
U+FA0C	≡ U+5140	FE40		(none)
U+5140	n/a	D8A3		
U+FA0D	≡ U+55C0	FE41		(none)
U+55C0	n/a	86D8		
U+FA18	≡ U+793C	FE47		(none)
U+793C	n/a	C0F1		
U+FA20	≡ U+8612	FE49		(none)
U+8162	n/a	CC55		

Annex D: Revision History

December 9, 2022: Added missing rows in table on page 5.