ISO/IEC JTC1/SC2/WG2 N 3691

Date: **2009-10-14**

ISO/IEC JTC1/SC2/WG2
Coded Character Set
Secretariat: Japan (JISC)

Doc. Type:	**Draft disposition of comments**
Title:	**Draft disposition of comments on SC2 N 4079 (ISO/IEC CD 10646, Information Technology – Universal Coded Character Set (UCS))**
Source:	**Michel Suignard (project editor)**
Project:	JTC1 02.10646.00.00.00.02
Status:	For review by WG2
Date:	2009-10-14
Distribution:	WG2
Reference:	SC2 N4079, 4088
Medium:	Paper, PDF file

Comments were received from India, Japan, Korea (ROK), U.K, and U.S.A. The following document is the draft disposition of those comments. The disposition is organized per country.

Note – The full content of the ballot comments have been included in this document to facilitate the reading. The dispositions are inserted in between these comments and are marked in **Underlined Bold Serif text**, *with explanatory text in italicized serif*.

India: Positive with comments

Technical comments

T1 Proposal to add one character in the Arabic block for representation of Kasmiri and annotation of existing characters

Kashmiri: Perso-Arabic script

The Kashmiri language is mainly written in the Perso-Arabic script. Arabic is already encoded in the Unicode to cater the requirement of Arabic based languages which includes Urdu, Kashmiri and Sindhi. Experts at Shrinagar University examined the Arabic Unicode for representation of Kashmiri language and opined that few characters need to be added for representation of the Kashmiri using Perso-Arabic script.

DIT organized meeting of experts to discuss the issues, which further examined the issues and recommended that one more character is required for addition in the existing Arabic code chart for representation of Kashmiri using Perso-Arabic script. The proposed character is shown in the table. In the Unicode for Arabic, code points Alef With Wavy Hamza Above and Alef With Wavy Hamza Below are encoded. Kashmiri uses Wavy Hamza Below with some other letters also. Hence there is need to encode this character in the Arabic block of the Unicode standard.

U+06xx Wavy Hamza Below

- Specifically used in other letters in Kashmiri

Experts also recommended that annotations need to be added in the existing Arabic codes for representation of Kashmiri using Perso-Arabic script.

06EA ARABIC EMPTY CENTRE LOW STOP

- Used in Kashmiri for palatalization

065A ARABIC VOWEL SIGN SMALL V ABOVE

- African languages
- Used in Kashmiri

06CC RABIC LETTER FARSI YEH

- Arabic, Persian, Urdu, Kashmiri
- Initial and medial forms of this letter have dots

→ 0649 arabic letter alef maksura
→ 064A arabic letter yeh

(see SC2 N4088 for further details)

Propose out of scope

Adding repertoire in the CD should be considered out of scope. This should be done through the amendment process. If there is a character of urgency for this character, it could be considered for Amendment 8. Otherwise it should be considered for future amendment.

Japan, Negative

Japan disapproves ISO/IEC CD10646 (SC2N4079) with comments below. Japan will change its vote if the comments are addressed appropriately.

JP.1 (Editorial): Title page

On the title page, the words "Multiple-Octet" are missing.

Propose non acceptance

The change is intentional. The standard is de-emphasizing the importance of stream encoding and clarifying the distinction between encoding scheme (serialization) and encoding form (how the abstract UCS value are representing in various forms). The abstract UCS values use a 6 digit integer form (000000 to 10FFFF). Given these changes, the multi-octet qualifier is misleading. These words have been removed from the title since at least WG2 N3509 (ISO/IEC 10646 working draft created in 2008).

JP.2 (Technical): On page 11, in 3 Normative references,

In Section 3, Normative references, a new material UAX#44 is added. Japan believes it is inappropriate and should be deleted. Also, the entire texts in 6.3.1 Classification on page 19 is only appropriate for the Unicode Standard (and its UAX#44) and it has no business with ISO/IEC 10646, so the entire subclause should be deleted.

Propose partial or non acceptance

Using property definitions to classify character type is a cornerstone of the new edition. It serves several purposes:

1) *Formally define character behavior that were before loosely defined in the standards such as Bidirectional behavior and Normalization.*
2) *Decrease the need to maintain enumeration of characters in the standard that easily get out of sync. The General Category value is a much more elegant way to describe character types (Graphic, Format, Control, etc…). In addition, it avoid errors as seen in the disposition of further comments.*

Before this proposed change, the standard was in fact deficient: the Unicode Bidi Algorithm and the Unicode Normalization Forms which have been normatively referenced for a long time, relies on property values specified on UAX#44.

However, it could be said that UAX#44 specifies more properties that necessary for ISO/IEC 10646. So it is possible to narrow down the normative references to these parts of UAX#44:
Following properties defined in UnicodeData.txt

- *General_Category (GC)*
- *Canonical_Combining_Class (used for normalization)*
- *Bidi_Class (used for Bidi algorithm)*
- *Decomposition_Type (used for normalization)*
- *Decomposition_Mapping (used for normalization)*
- *Bidi_Mirroring (used for Bidi algorithm)*

All properties defined in DerivedNormalizationProps.txt (total of 12 used for Normalization).

This would, however, make the reading of clause 3 much more complex. It is not clear there is any harm in creating normative references to a superset of properties. The key goal is to ensure that properties that are shared between the Unicode Standard and ISO/IEC 10646 are specified once. This is the only way to make sure the two standards stay symchronized.

JP.3 (Technical): On page 11, 4.1, Base Character

The definition should read "A graphic character that does not graphically combine with preceding characters."

Propose non acceptance

The former definition (requested here by Japan) is in fact imprecise. It does not give implementers any formal way to determine the repertoire subset concerned by this type. The same could be said for the Combining characters. The new edition by using GC types creates an exact definition of these categories.

JP.4 (Technical): On page 11, 4.4, Canonical Representation

"canonical representation" should read "Canonical form" as in the previous editions.

Propose non acceptance

A major clarification of this edition is to separate 3 concepts:

- *Canonical representation which is how character are represented in the UCS code space,*
- *Encoding forms which represent the various forms on which a character can be encoded*
- *Encoding schemes which define how to serialize encoding forms.*

The previous edition was muddling all these concepts together. There is no such thing as a canonical form, unless you want to give a preference to the UCS-4 encoding form against the other encoding forms. While researching this issue, the editor found an inconsistency in the 2nd paragraph of clause 5 General structure of the UCS, where 'canonical form' should be replaced with 'canonical representation'.

JP.5 (Technical): On page 13, 4.18, Control Character

The definition should read "A control function the coded representation of which consist of a single code position".

Propose acceptance in principle

The definition will read "A control function the coded representation of which consists of a single code point". The term 'code point' is preferred, to better synchronize terminology between Unicode and 10646.

JP.6 (Technical): On page 16, 4.58, Unpaired surrogate code unit

The term "Unpaired surrogate code unit" substituting "RC-element" in the previous editions seems too verbose. Some shorter words, e.g., just "unpaired surrogate", is better.

Propose non acceptance

Strictly speaking "Unpaired surrogate code unit' is replacing "Unpaired RC-element". The term 'surrogate' by itself is ambiguous because it can either represent a pair of code units as in 'surrogate pair' or a single code unit when it is unpaired. Therefore the definition as stated in the CD should stay.

JP.7 (Editorial): On page 30, in 16.3 Format Characters,

The character 10A3F KHAROSHTHI VIRAMA is missing from the list.

Propose non acceptance

In the previous edition, there were many characters that were mistakenly assumed to be format characters, mostly based on the fact that their representative glyphs was using a dotted square. By using formal reference to the General_Category property, these deficiencies have been corrected. While some characters have been removed from the list (such as 10A3F KHAROSHTHI VIRAMA and more), some others have been added (such as 17B4 KHMER VOWEL INHERENT AQ). It should be noted that some format characters are not described in Annex F. Although the annex is only informative, they should probably be described there:

17B4	*KHMER VOWEL INHERENT AQ*
17B5	*KHMER VOWEL INHERENT AA*
1A60	*LANNA TAI THAM SIGN SAKOT*
1CBF	*MEITEI MAYEK SIGN VIRAMA*
2061	*FUNCTION APPLICATION*
2062	*INVISIBLE TIMES*
2063	*INVISIBLE SEPARATOR*
2064	*INVISIBLE PLUS*

JP.9 (Technical): On page 37-38, clause 23.1 List of source references

Clause 23.1 defines the abbreviated names of the source references for CJK Unified Ideographs. Most names are used in the syntax to indicate the sources in CJKU_SR.txt which is described in clause 23.2 (on page 40), but some names are different. For example, Chinese source Kangxi Dictionary ideographs are abbreviated as "G_KX" in clause 23.1, and abbreviated as "KX" in clause 23.1 and CJKU_SR.txt. ISO/IEC 10646:2003 used "G_KX" only.

In addition, some abbreviations in the code charts are different from clause 23.1 and 23.2. For example, U+2A701 is noted as "G_ZJW00001" in CJKU_SR.txt, and noted as "GZJW00001" in the code chart (p. 2040). This difference was not found in the code chart included in ISO/IEC 10646:2003/Amd.5:2008.

Using same abbreviations is better.

WG2 discussion

See also Korean comment T.2.

The addition of source references in the chart has created new challenges, namely their length which is sometimes too large to fit in the available space in the cell. This has resulted in unintended differences between clause 23.1, clause 23.2, and the charts. This concerns mostly G sources. In most cases, the underscore has been removed from the G source names. In one case: G_KX, because the preferred format for charts has been KXdddd.dd, and the two characters 'G_' were removed. With that change it could look like a Korean source, although KX is a well know acronym for KangXi.

This discrepancy also concerns one Japanese source: J_ARIB which was shortened to JARIB. It is trivial to fix 23.1 and 23.2 to represent exactly what is in the chart. However the editor would like to get some guidance, given all the production constraints. Note that the problem is more acute in the charts used for ext A, B, C, and D, because the source value is the only way to determine the IRG source.

There are several issues with G source references as they stand today for use in charts or even as references:

- *Their prefix is sometimes too long and redundant; for example is the 2nd 'G' in G_GFHZBddddd' necessary?*
- *Some entries are duplicate ones, such as 'G_HZ' and 'G_HZddddd'; all such entries should use the 'G_HZddddd'.*
- *The entries with no numerical value attached are basically useless; they should be augmented with a numerical value. Except for 'G-4K' and 'G_CY', all others references have already a variant with numerical values, so it is simply a matter of providing the information to the editor.*

Concerning the KX entry, it is true that the 'K' prefix may be confusing, so we should probably add back the 'G' prefix. Note that we may have to remove the period between the page and the index into the page if the index does not fit in the cell.

Following is a proposal for new identifiers:

Old 23.1	New 23.1	Old 23.2	New 23.2	Comment
G0	G0	G0-hhhh	G0-hhhh	*Unchanged*
G1	G1	G1-hhhh	G1-hhhh	*Unchanged*
G3	G3	G3-hhhh	G3-hhhh	*Unchanged*
G5	G5	G5-hhhh	G5-hhhh	*Unchanged*
G7	G7	G7-hhhh	G7-hhhh	*Unchanged*
GS	GS	GS-hhhh	GS-hhhh	*Unchanged*
G8	G8	G8-hhhh	G8-hhhh	*Unchanged*
G9	G9	G9-hhhh	G9-hhhh	*Unchanged*
GE	GE	GE-hhhh	GE-hhhh	*Unchanged*
G_4K	G4K	G_4K	G4K????	*Need to determine index range and base*
G_BK	GBK	G_BK or G_BKddddd	GBKddddd	
G_CH	GCH	G_CH or G_CHddddd	GCHddddd	
G_CY	GCY	G_CY	GCY????	*Need to determine index range and base*
G_CYY	GCYY	G_CYYddddd	GCYddddd	
G_FZ	GFZ	G_FZ or G_FZddddd	GFZddddd	
G_GFHZB	GFHZB	G_GFHZBddd	GFHZBddd	*One 'G' removed*
G_GH	GH	G_GHddddd	GHdddddd	*One 'G' removed, Old notation had a digit missing*
G_GJZ	GJZ	G_GJZddddd	GJZddddd	*One 'G' removed*
G_HC	GHC	G_HC or G_HCddddd	GHCddddd	
G_HZ	GHZ	G_HZ or G_HZddddd	GHZddddd	
G_IDC	GIDC	G_IDCddd	GIDCddd	
G_XC	GXC	G_XCddddd	GXCddddd	
G_ZFY	GZFY	G_ZFYddddd	GZFYddddd	
G_ZJW	GZJW	G_ZJWddddd	GZJWddddd	
G_KX	GKX	KXdddd.dd	GKXdddd.dd	*The dot may be removed if not enough space*

All these formats are already in use for the chart, except for G4K????, GCY????, and GKXdddd.dd.

In addition, it could be interesting to merge CJKC_SR.txt and CJKU_SR.txt that have almost the same format and the related Unicode data file now that Unihan has been split in many files. This would simplify greatly the synchronization between the two standards.

JP.10 (Technical): On page 39 in Clause 23.1 List of source references,

The definition for JH source should be as follows:

JH Hanyo-Denshi Program (汎用電子情報交換環境整備プログラム)

Propose acceptance

Also requested for PDAM8

JP.11 (Editorial): On page 39, in Clause 23.1 List of source references,

J_ARIB source should be corrected as JARIB.

Propose acceptance in principle

See disposition of comment JP.9.

JP.12 (General): On page 47, in 25 Named UCS Sequence Identifiers.

The current text does not show any actual named UCS sequence identifiers and says "specified in the Unicode Standard UAX#34". Japan considers it is inappropriate. It makes UCS-Unicode synchronization more difficult. In particular, the current version of UAX#34 doesn't include new named USIs added in Amd.7. The international standard should carry its own list of named USIs. For maintenance reasons, it may be a good idea to isolate the list from the main content and to create a new annex for the list of NUSIs.

UAX#34 should be deleted from the list of normative references on page 11.

Propose acceptance in principle

The editor does not see how it could be more difficult to synchronize if the content is defined in a single place. However, the Japanese raises a valid issue concerning the timing of UAX#34 updates. A possible solution is to create a linked file from clause 25 which would have the same format as the file linked to by UAX#34. This would remove the need to reference normatively UAX#34.

JP.13 (Editorial): On page 53 in 30.2 Character names list, 3rd line,

"code position" should read "code point", since WG2 decided to prefer the latter term than the former.

Accepted

JP.14 (General): J column font

J column font for U+20B9F (part of CJK UNIFIED IDEOGRAPHS EXTENSION B) and for CJK UNIFIED IDEOGRAPHS EXTENSION D need appropriate update.

Propose acceptance in principle

The character 20B9F was not part of the font package received from the Japanese NB. The column will be updated as soon as the editor gets a set of new Japanese fonts. The J column font for Ext D will be fixed as a result of accepting the relevant Japanese comments on PDAM8.

JP.15 (Editorial): On page 2182 in Table I.1,

"IDS example represents" column is broken.

Accepted

JP.16 (General): On page 2194 in Annex P.

The title of the annex in this draft is changed from the previous editions and contains the information on CJK Unified Ideographs only (as in the new title.) Information for other characters are removed from the Annex P. Japan understands the intention of this change was that the new Character Name List contains equivalent information and additional texts in Annex P were considered redundant. However, some information that were available in the previous editions are missing in the draft.

For example, in the previous editions, the following text was included in the Annex P:

0218 LATIN CAPITAL LETTER S WITH COMMA BELOW

This character is intended for use only in those cases where it is necessary to make a distinction from the letter with cedilla. Both forms of the letter may be found in a single document written in a single language, e.g. Romanian or Turkish. In ISO/IEC 8859-2 only a single (8-bit) coded character is provided, LATIN CAPITAL LETTER S WITH CEDILLA, which maps to 015E in ISO/IEC 10646 by default, and may map by mutual agreement between sender and receiver to this letter with comma below. See ISO/IEC 8859-2 for further information on the use of that standard.

On the other hand, the Character name list of the current draft doesn't contain any information regarding characters with comma below v.s. with cedilla.
Japan considers this type of information formerly available in Annex P is important and should be included in the next edition of the standard. One way to do so is to revert to the previous edition's Annex P.

Propose acceptance in principle

When the new chart format was adopted, a consequence was that a lot of information was added in the standard through the various chart new fields. This made the need for Annex P mostly irrelevant. The editor worked with the Unicode editorial committee in order to add information which was part of Annex P and missing in the new chart format. In that particular case there is indeed information concerning the comma below versus cedilla but in an abbreviated form. In addition, in the chart the information is only provided for the lower case form of a character. Typically, the chart information is expressed in a format terser than the original Annex P. For example the new information for the example shown by Japan is as following:

015E LATIN CAPITAL LETTER S WITH CEDILLA
≡ 0053 0327
015F LATIN SMALL LETTER S WITH CEDILLA
** Turkish, Azerbaijani, ...*
** the character 0219 is preferred for Romanian*
→ 0219 latin small letter s with comma below
≡ 0073 0327
0162 LATIN CAPITAL LETTER T WITH CEDILLA
≡ 0054 0327
0163 LATIN SMALL LETTER T WITH CEDILLA
** Semitic transliteration, ...*
** the character 021B is preferred for Romanian*
→ 021B latin small letter t with comma below
≡ 0074 0327
0218 LATIN CAPITAL LETTER S WITH COMMA BELOW
≡0053 0326
0219 LATIN SMALL LETTER S WITH COMMA BELOW
** Romanian*
→ 015F latin small letter s with cedilla
≡ 0073 0326
021A LATIN CAPITAL LETTER T WITH COMMA BELOW
≡ 0054 0326
021B LATIN SMALL LETTER T WITH COMMA BELOW
** Romanian*
→ 0163 latin small letter t with cedilla
≡ 0074 0326

This is probably one of the rare cases where the former Annex P contained more information than the chart pages. One suggestion is to augment the annotation for that list as following (new text highlighted):

015E LATIN CAPITAL LETTER S WITH CEDILLA
** map to ISO/IEC 8859-2 latin capital letter s with cedilla*
≡ 0053 0327
015F LATIN SMALL LETTER S WITH CEDILLA
** Turkish, Azerbaijani, ...*
** map to ISO/IEC 8859-2 latin small letter s with cedilla*
** the character 0219 is preferred for Romanian*
→ 0219 latin small letter s with comma below
≡ 0073 0327
0162 LATIN CAPITAL LETTER T WITH CEDILLA
** map to ISO/IEC 8859-2 latin capital letter t with cedilla*
≡ 0054 0327
0163 LATIN SMALL LETTER T WITH CEDILLA
** Semitic transliteration, ...*

** map to ISO/IEC 8859-2 latin small letter t with cedilla*
** the character 021B is preferred for Romanian*
→ 021B latin small letter t with comma below
≡ 0074 0327
0218 LATIN CAPITAL LETTER S WITH COMMA BELOW
≡0053 0326
0219 LATIN SMALL LETTER S WITH COMMA BELOW
** Romanian*
** used when distinct comma below form is required*
→ 015F latin small letter s with cedilla
≡ 0073 0326
021A LATIN CAPITAL LETTER T WITH COMMA BELOW
≡ 0054 0326
021B LATIN SMALL LETTER T WITH COMMA BELOW
** Romanian*
** used when distinct comma below form is required*
→ 0163 latin small letter t with cedilla
≡ 0074 0326

JP.17 (Editorial): On page 2153 in A.3.2 299 BMP FIRST EDITION,
The last line, "IO/IEC 10646-1" should read "ISO/IEC 10646-1".
Accepted

JP.18 (Editorial): On page 2160, in A.6.2 304 UNICODE 3.2,
The first line, "from A.1 and A.1 and several ranges" should read "from A.1 and several ranges".
Accepted

JP.19 (General): On page 2178 in Annex G,
The file name “Allnames.txt” in NOTE 1 has been amended at every amendment. Has this been done by design?
Noted
Yes, each amendment requests that a new file called Amxnames.txt (x standing for the amendment number) be inserted ‘in the list of character names in Annex G’. Note that the actual result is not provided in the amendments.

JP.20 (Editorial): On page 2185 in Annex L, Guideline 2, NOTE,
The words “as for the control characters” in the previous editions has been changed to “as control characters”. Is this intentional?
WG2 discussion
That note said in the previous edition:
NOTE – In ISO/IEC 6429, also the names of the modes have been presented in the same way as control functions.
Now it says:
NOTE – In ISO/IEC 6429, the names of the modes have also been presented in the same way as control functions.
(note the move of the adverb ‘also’)
Neither the previous nor the new edition use the term ‘control characters’.

JP.21 (Editorial): On page 2188 in Annex M,
The vertical space before the first line starting with the words “ISO 233:1984” is too small.
Accepted

JP.22 (Editorial): On page 2193 in N.3,
The items ”BMP-form (2)” and ”ISO 10646 form 2” are missing.
Propose partial acceptance
It is the intent of this new edition to deprecate these encoding forms. It is mentioned in a note in clause 9.2 and the Note 2 in N.3. An additional note concerning the related object descriptor could be added at the end of N.3 as follows:

NOTE 3 – Previous versions of this standard supported the “ISO 10646 form 2” object descriptor which is now deprecated.

JP.23 (Editorial): On page 2195 in Annex Q, the second line in the NOTE,

The expression "code .point" should read "code point" (no period before "point".)

Accepted

JP.24 (General): On page 2195, Annex Q

Annex Q in this draft contains nothing useful. Japan wants to see opinions from the representing bodies of the major user community of the script (i.e., KR and KP) before removing the mapping table from the standard. If it is ok to remove the table, Japan sees no purpose of this annex and prefers to remove Annex Q entirely (as opposed to leave it as an empty annex with one NOTE as in the current draft.)

Propose noted

The only reason this annex is maintained is to keep the numbering sequences of the following annexes. There are important annexes following, such as Annex S, which are widely known to represent a specific topic and it would be very disturbing to the user community to have the annexes numbers changing. Creating place holder for deprecated annexes is a principle commonly used in ISO standard revision. Annex Q was referring to a state of the standard which has been deprecated for a long time and does not need to be preserved.

JP.25 (General): On page 2196, Annex R,

The formatted table versions (PDF) that were available in the previous editions are removed in this draft. Japan wants to see opinions on this change from the representing bodies of the major user community of the script (i.e., KR and KP).

Propose noted

Note that the non formatted version of the Hangul syllable names (also part of annex R) has been preserved. This is technically equivalent to the formatted version and is much more usable by implementers and users of the standard. No comments from other NBs have been received on this topic.

JP.26 (Editorial): On page 2197, in Annex S,

First line, "30" should read "clause 30".

Accepted

JP.27 (Technical): On page 2199-2200, in S.1.4,

Several glyph shape examples have been removed or modified from the previous editions:

In "b)", Two glyph shapes of the character 告 have been deleted.

In "c)", Two glyph shapes of the character 児 have been deleted. Two glyph shapes of the character奔 have been deleted also.

In "e)", Two glyph shapes of the character 西 have been deleted.

In "f)", The two glyph shapes for each of the characters 父 and 丈 are swapped and reordered. Two glyph shapes with the kanji radical 廴 have been deleted

Propose noted

The new edition has modified annex S to use far less images than in the previous editions. Many glyphs for the CJK examples are not easily available in any fonts. The editor is open to recreate a version closer to the original if fonts are provided to create these examples. In the meantime, it is the editor's belief that the purpose of the examples has been preserved in the new edition.

JP.28 (Technical): On page 2200, in S.2.1,

The Edition of the Kangxi Dictionary has been changed from 7th edition to 9th edition. Japan is unsure why. If it is an editorial mistake, change it to 7th edition.

Also, the word "Beijing" is part of edition, not part of the title of the book, so the table format is inappropriate.

Accepted

This was the result of a copy/paste error when creating the new document. The intent is as suggested by Japan.

JP.29 (General): code chart for CJK Unified Ideographs

Japan believes that there are too many editorial errors and quality problems in the new code chart for CJK Unified Ideographs in the CD. Following are major issues found in the the CD ballot review. More problems may be found by further checking. Japan is negative to include the new code chart in the next edition of 10646.

<u>WG2 discussion</u>

It is true that there are a number of significant issues in the charts for CJK Unified Ideographs. Many are due to the lack of available fonts and should be solved before the next edition is published through the supply of these fonts to the editor. Some issues mentioned here already existed in the previous edition (such as lack of KP_source glyphs). There is no solution involving the usage of 'old code charts' that could be entertained for creating a new edition. Therefore, either we fix the major issues in the CD in timely fashion, or we delay the publication. At the same time, given the large size of the repertoire, perfection is only a goal.

a) KP-source glyphs are entirely missing.

<u>WG2 discussion</u>

Except for Ext C, KP-source glyphs have never been provided. So this is not a regression. We have to either keep it blank and explain why, or remove the column altogether.

b) Glyph of K4-0005 (U+200CE) is missing.

<u>Propose acceptance</u>

Further examination of the font provided by South Korea (CJK_ExtB_KR-v40) shows that the font has a defect for that character (glyph available but not mapped). The editor can either fix the font or get a new font from Korea.

c) Glyph of J3-7A6D(u+08362) is wrong.

(JNB verified the font it submitted to the project editor was correct.)

<u>Propose acceptance</u>

Japan submitted 2 fonts: RgHeiseiM-W3 and JapanBMPS to cover that block and Extension A.

The character U+8362 is displayed as 荢 *with RgHeiseiM-W3 and* 荢 *with JapanBMPS. The editor is therefore assuming that the glyph provided by the JapanBMPS is preferred in that occurrence. As of now, the chart program takes in cascade the two fonts, with RgHeiseiM-W3looked up first. This character will have a special case built to make sure JapanBMPS is selected for that character.*

d) serious problems on unification criteria of CJK Unified Ideographs

The changes of following glyphs in the new fonts caused serious problems on unification criteria of CJK Unified Ideographs.

<u>WG2 discussion</u>

For the discussion below, the editor has added the glyphs as available for chart production; the reference glyph is taken from the G column (in absence of clearer source) or T column for ext C. It should also be noted that the editor have only received a subset of the needed fonts from some NBs. This explains all errors concerning the T and V sources. Details following.

V1-652F(u+083eb) 菫 and u+05807 堇 have the same shapes.

The V source glyph for 83EB is in error, it looks different from the glyph used in ISO/IEC 10646:2003

K2-2C22(u+059ba) 妺 and u+59b9 妹 have the same shapes.

In the original chart from ISO/IEC 10646:2003, the glyphs also look the same. These two characters were dis-unified because of the source separation rules for the G source. It isn't clear whether or not other IRG sources, not part of the list mentioned in S.1.6 can the use the dual mapping or not. In all cases, this does not look like a regression.

T2-565F(u+05b14) 嬔 and u+05b0e 嬎 have the same shapes.

The editor did not get a proper font for the main CJK block from Taiwan. The commercial font used instead for U+5B14 (MingLiu) is in error, which could be fixed by swapping its glyph for 5B0E and 5B14.

T3-4B5F(u+05b0e) 嬎 and u+05b14 嬔 have the same shapes.

Same issue and resolution as character above

V1-5669(u+062d4) 拔 and u+062e4 拤 have the same shapes.

They clearly don't have the same shape, probably a typo in the comment?

V1-5B47(u+06bbc) 殼 and u+06bbb 殻 have the same shapes.

The V source glyph for 6BBC is in error, it looks different from the glyph used in ISO/IEC 10646:2003

V1-5B49(u+06bc0) 毀 and u+06bc1 毁 have the same shapes.

The V source glyph for 6BC0 is in error, it looks different from the glyph used in ISO/IEC 10646:2003

V1-6172(u+07b53) 筓 and u+07b04 笄 have the same shapes.

The V source glyph for 7B53 is in error, it looks different from the glyph used in ISO/IEC 10646:2003

V2-8E80(u+07ca4) 粤 and u+07cb5 粵 have the same shapes.

The V source for 7CA4 already existed in the source reference file for ISO/IEC 10646:2003 but was not shown in the chart. As of now the V source glyph for 7CA4 seems to be in error.

T5-2150(u+20506) 𠔆 and u+20505 𠔅 have the same shapes.

The editor did not get a proper font for Ext B from Taiwan. The commercial font used instead (MingLiu-extB) for U+20506 is in error.

V1-4E3E(u+053f1) 叱 and u+20b9f 𠮟 have the same shapes.

The shapes are different, but the V source for U+53F1 could be drawn better.

T2-6D4B(u+05dd5) 巕 and u+21fd2 𡿒 have the same shapes.

The editor did not get a proper font for the main CJK block from Taiwan. The commercial font used instead for U+5DD5 (MingLiu) is in error.

K3-2A34(u+03dd7) 㷗 and u+242c5 𤋅 have the same shapes.

The glyph part of CJK_BMP_KR-v40 is in error. It is also different from the K source glyph for U+3DD7 in ISO/IEC 10646:2003.

V2-8836(u+2a01a) 𪀚 and u+29fe0 𩿠 have the same shapes.

The V source glyph for 2A01A seems to be in error, not provided for ISO/IEC 10646:2003.

V1-575F(u+06492) 撒 and u+2abab 𪮫 have the same shapes.

The V source glyph for 6492 is in error, it looks different from the glyph used in ISO/IEC 10646:2003

V0-423F(u+08326) 茦 and u+2b1ff 𫇿 have the same shapes.

The V source glyph for 8326 is in error, it looks different from the glyph used in ISO/IEC 10646:2003

e) Following characters have significantly different shape from the current standard.
We need to make sure the shapes in the draft are correct. Japan wants input from IRG for the point.
(glyphs added by editor according to chart fonts)

T3-406C(u+03712) 㜒

KX0270.25(u+0371c) 㜜

KX0271.37(u+03729) 㜩

GS-605E(u+04599) 䖙

G3-7827(u+04c57) 䱗

V1-4C22(u+050b7) 傷

T2-4C61(u+05284) 劄

T3-3B55(u+060d6) 惖

V2-8C30(u+06176) 慶

T4-4452(u+06f45) 潅

V2-8D5B(u+06f75) 潵

GE-3137(u+06ff2) 濲

V1-5F3A(u+074ca) 瓊

V2-8E47(u+07a37) 稷

V3-3664(u+07e06) 縆

V1-6473(u+08319) 茙

GE-3B73,K1-5B7D(u+08641) 虁 虁

V1-6C61(u+09b2d) 鬭

V2-6E38(u+200ab) 𠂫

V0-3023(u+20133) 𠄳

V2-6E4A(u+2020b) 𠈋

T5-2526(u+20215) 𠈕

V0-4422(u+204af) 𠒯

T6-353E(u+2051c) 𠔜

V2-6E7D(u+2052e) 𠔮

T5-6135(u+2061e) 𠘞

V2-6F47(u+207e4) 𠟤

V0-3053(u+207f2) 𠟲

V2-6F5D(u+20927) 𠤧

V0-4574(u+20a21) 𠨡

V0-314B(u+20CD0) 𠳐

V0-3361(u+21088) 𡂈

T7-4724(u+2292a) 𢤪

V2-7671(u+22b31) 𢬱

V0-3A52(u+23659) 𣙙

T5-6223(u+23758) 𣝘

T7-4749(u+2379b) 𣞛

T5-555F(u+24b88) 𤮈

T5-4770(u+253f0) 𥏰

V3-3739(u+26bec) 𦯬

T7-5F67(u+283ec) 𨏬

T4-4832(u+290aa) 𩂪

TD-5B55(u+2aac9) 𪫉

TE-353F(u+2ad12) 𪴒

WG2 discussion

The editor corrected the entry ‘C3-3664(u+07e06)’ to V3-3664(u+07e06). As mentioned before, some of these entries (from T and V sources) were not provided by NBs and as such may have errors. Among the entries provided above, the following items correspond to fonts provided by NBs.

KX0270.25(u+0371c) 㜜

KX0271.37(u+03729) 㜩

GS-605E(u+04599) 䖙

G3-7827(u+04c57) 䱗

GE-3137(u+06ff2) 濲

GE-3B73,K1-5B7D(u+08641) 虁 虁

TD-5B55(u+2aac9) 𪫉

TE-353F(u+2ad12) 𪴒

f) Following characters have the source references of KangXi dictionary,
but their new shapes look different from that of KangXi dictionary.

KX0956.36(u+04389) 䎉

KX1030.13(u+044bb) 䒻

KX1319.05(u+04949) 䥉

WG2 discussion

No opinion from the editor.

Korea (ROK): Negative

Technical comments:

T1. p. 39, 23.1 List of source references

[current text]
The Hanja K sources are
K0 KS C 5601-1987
K1 KS C 5657-1991
K2 PKS C 5700-1 1994
K3 PKS C 5700-2 1994
K4 PKS 5700-3:1998
K5 Korean IRG Hanja Character Set 5th Edition: 2001
--->
[proposed text]
The Hanja K sources are
K0 KS X 1001:2004 (formerly KS C 5601)
K1 KS X 1002:2001 (formerly KS C 5657)
K2 PKS X 1027-1
K3 PKS X 1027-2
K4 PKS X 1027-3
K5 PKS X 1027-4

WG2 discussion

The editor is not against revising source references, but because it touches the identity of characters with such references, it creates stability concerns. Identity is established by the reference as it existed when it was used. If the reference is updated, it may not be necessary usable as a reference for existing characters. Unless the new source reference is a mere repackaging of the former version, it is not wise to do so. This is why many of the IRG source references are still dated from the 1990, even if more recent versions exist.
From access to other reference material such as 'CJKV Information Processing, by Ken Lunde, 2nd edition), and after discussion with Ken Lunde, the editor would propose to add the former references to all 6 entries as follows (Years added for the first two entries as compared to the Korean NB suggestion):
K0 KS X 1001:2004 (formerly KS C 5601-1987)
K1 KS X 1002:2001 (formerly KS C 5657-1991)
K2 PKS X 1027-1 (formerly PKS C 5700-1 1994)
K3 PKS X 1027-2 (formerly PKS C 5700-2 1994)
K4 PKS X 1027-3 (formerly PKS 5700-3:1998)
K5 PKS X 1027-4 (formerly Korean IRG Hanja Character Set 5th Edition: 2001)

T2. p.40, 23.2, third bullet

3rd field: Hanzi G sources(G0-hhhh), (G1-hhhh), (G3-hhhh), (G5-hhhh), (G7-hhhh), (GS-hhhh), (G8-hhhh), (G9-hhhh), (GE-hhhh), (G_4K), (G_BK), (G_BKddddd), (G_CH), (G_CY), (G_CYYddddd), (G_CHddddd), (G_FZ), (G_FZddddd), (G_GHddddd), (G_GFHZBddd), (G_GJZddddd), (G_HC), (G_HCddddd), (G_HZ), (G_HZddddd), (G_IDCddd), (G_XCddddd), (G_ZFYddddd), (G_ZJWddddd), or (KXdddd.dd)
[current text] KXdddd.dd
→
[proposed text] GKXdddd.dd (or G_Kdddd.dd or G_Xdddd.dd or G_KXdddd.dd)
Rationale: Since all G sources except KX start with 'G', we propose to change KXdddd.dd to GKXdddd.dd (or G_Kdddd.dd or G_Xdddd.dd, or G_KXdddd.dd) so that all G sources start with 'G'. Furthermore, considering that all Hanja K sources starts with 'K' and and Hanja KP sources starts with 'KP', 'G_K...' or 'G_X...' look much better than 'KX...'

WG2 discussion

See disposition of comment JP9 from Japan. That disposition proposes to use 'GKXdddd.dd'.

T3. p.41, 23.3, third para.

The code chart for the CJK UNIFIED IDEOGRAPHS block (4E00-9FFF) uses a fixed column format (i.e. source references from a given source always appear in the same column) while the code charts for the other CJK Unified blocks show graphic symbols per the following order of appearance: G, T, J, K, V, KP, H, U, and M
[current text] G, T, J, K, V, KP, H, U, and M
→
[proposed text] G, T, J, K, KP, V, H, M, and U
Rationale:
1) KP precedes V (e.g., p. 388, U340C)
2) In CJKU main, M (in the second column) precedes U in the third column).
Note. In CJKU Extensions, there is no Hanja char having both M and U source references.

Propose acceptance

In fact, the order described in 23.3 (G, T, J, K, V, KP, H, U, and M) is not used in the chart, as observed by Korea for character 340C. The order is the same for all blocks, with the main block having holes for unused references. Therefore the proposed text 'G, T, J, K, KP, V, H, M, and U' is correct.

T4. p. 42, 23.3.2 *Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION A*

The following figure shows the presentation for the CJK UNIFIED IDEOGRAPH EXTENSION A block. Up to four sources per characters are represented in a single row. If more than four sources exist, an additional row is used.
[current text] ... four sources ...
→
[proposed text] ... three sources ... [occurs twice]

Accepted

For a while it was considered possible to fit four sources per row. This proved impossible at production time.
See also comment E8 from UK NB.

T5. p. 43

1) There is no explanation as to which country will provide font for CJK Compatibility characters shared by more than one country. We suggest to discuss and to add an explanation to the Standard.
For your information, the sharing status is as follows:
KJ 6, KPJ 1, KH 1, PT 49, PTH 1, TH 4, 62 in total (out of 1,000 CJKC chars)
2) Rep. of Korea will provide the font for U0F900 ~ U0FA0B and requests that the font be used for printing U0F900 ~ U0FA0B.

WG2 discussion

Today, compatibility CJK characters are shown in the chart with a single glyph per character. It would be desirable to allow multi-column display for these characters as well for at least two reasons:
Some CJK compatibility characters are shared (as mentioned by Korea)
Some characters in the CJK Compatibility Ideograph block are in CJK Unified Ideographs and should be treated as the other unified characters.
This would require an update to the chart program and a new clause after 23.4 describing the source reference presentation for CJK Compatibility ideograph blocks.
See disposition of comment T3 from USA NB for further consideration on this topic.

T6. p. 53

[current text]
Decompositions, preceded by '≡', or '≈', describing various mapping between characters.
→
[proposed change] There seems no explanation as to usage difference of these two characters'≡','≈.
1) As a result, we could not review properly lines having these characters. We need to review such lines "after" explanations are given.
2) We suggest that explanations about the usage difference be added.

WG2 discussion

That text was introduced by ISO/IEC 10646:2003 amendment 5 along with the new chart format. Documenting '≡' and '≈' requires introducing concepts that were simply hinted in ISO/IEC10646 (in clause 21 of this CD) but are fully explained in the Unicode Standard. The following is a proposal to introduce some term and definitions. They *should also make easier to read related text in the Unicode Standard components which are normatively referenced (such as UAX#15 Normalization forms).*

Two options are possible, either remove from the standard description using terms that are not fully integrated in the standard, or extend the standard to add these formal definitions.

Option 1 (simplification)

Modification of existing terms in clause 4 Terms and definitions

Canonical representation

The representation with which each character of this coded character set is specified using a single code point within the UCS codespace.

Introduction of a new term in clause 4 Terms and definitions

Decomposition mapping

A mapping from a character to a sequence of one or more characters that is a canonical or compatibility equivalent.

In clause 21 Normalization forms, replace the four items list as follows:

1) Normalization Form D (NFD),
2) Normalization Form C (NFC),
3) Normalization Form KD (NFKD),
4) Normalization Form KC (NFKC).

In clause 30.2 Character name names list, replace the last informative item by the following

- Decomposition mapping, preceded by '≡' for canonical mapping, and by '≈' for compatibility mapping.

Option2 (extension)

Modification of existing terms in clause 4 Terms and definitions

Canonical representation

The representation with which each character of this coded character set is specified using a single code point within the UCS codespace.

Introduction of new terms in clause 4 Terms and definitions

Decomposition mapping

A mapping from a character to a sequence of one or more characters that is a canonical or compatibility equivalent. The mapping is specified for each character by the Decomposition_Mapping property as defined in the Unicode Character Database (see 3). The mapping may contain formatting information.

Compatibility mapping

Decomposition mapping for a character specified either by its Decomposition_Mapping property value when it contains formatting information, or by itself.

Canonical mapping

Decomposition mapping for a character specified either by its Decomposition_Mapping property value when it contains no formatting information, or by itself.

Canonical decomposition

Decomposition of a character or character sequence that results from recursively applying the canonical mappings until no characters can be further decomposed, and then reordering combining characters according to their canonical ordering.

Compatibility decomposition

Decomposition of a character or character sequence that results from recursively applying both the compatibility mappings and the canonical mappings until no characters can be further decomposed, and then reordering combining characters according to their canonical ordering.

Canonical ordering

Process by which combining characters are ordered in the increasing value of their combining class.

In clause 21 Normalization forms, add the following paragraph after the notes.

Canonical composition is a term defined in the Unicode standard UAX#15 as part of that specification.

In clause 30.2 Character name names list, add the following items to the normative list, along with a new note:

- Canonical mapping, preceded by ‘≡’ , when the canonical mapping is not the character itself,
- Compatibility mapping, preceded by ‘≈’, when the compatibility mapping is not the character itself. It also may contain an informative formatting tag between brackets.

NOTE – The following formatting tags corresponding to formatting information found in the Unicode Character Database (see 3) are used:

<font>	A font variant (for example, a blackletter form)
<noBreak>	A no-break version of a space, hyphen, or other punctuation
<initial>	An initial presentation form (Arabic)
<medial>	A medial presentation form (Arabic)
<final>	A final presentation form (Arabic)
<isolated>	An isolated presentation form (Arabic)
<circle>	An encircled form
<super>	A superscript form
<sub>	A subscript form
<vertical>	A vertical layout presentation form
<wide>	A fullwidth compatibility character
<narrow>	A halfwidth compatibility character
<small>	A small variant form
<square>	A square font variant
<fraction>	A vulgar fraction form
<compat>	Otherwise unspecified compatibility character

In the character names list, the <compat> label is suppressed.

In the same clause, remove the last item in the informative list (starting with ‘Decompositions, preceded’).
-end of Option 2

The changes in these 2 options do not change in substance the technical content of the standard.
In option 1), the formal reference to UAX#15 takes care of the details related to mappings, and any details surfacing in the standard, such as the chart name list, stay informative.
In option 2), more terms and definitions are added, which makes the mapping content normative to ISO/IEC 10646.
In different ways they clarify the roles of the compatibility and canonical mappings shown in the chart. As a consequence, changing the ‘≈’notation in the chart pages representing compatibility mapping would break the stability of compatibility normalization forms.

T7. p. 369, left column, top

[current text]
3131 ㄱ HANGUL LETTER KIYEOK
≈1100 ㄱ hangul choseong kiyeok
→
[proposed text]
3131 ㄱ HANGUL LETTER KIYEOK
→ 1100 ㄱ Hangul choseong kiyeok
≈ FFA1 ㄱ Halfwidth hangul letter kiyeok

Rationale: The usage of U3131 and UFFA1 is fairly similar. In contrast, the usage of U1100 is quite different from that of U3131 and UFFA1.
= Similar changes are proposed for code positions U3132 ~ U3163.

Propose non acceptance

See disposition of comment T6.
We have now:
3131 ㄱ HANGUL LETTER KIYEOK
≈ 1100 ㄱ hangul choseong kiyeok
And
FFA1 ㄱ HALFWIDTH HANGUL LETTER KIYEOK
≈ <narrow> 3131 ㄱ

Changing the first compatibility mapping without changing the second would result in a loop when compatibility decomposition is applied. As of now, compatibility decomposition for 3131 is a single step to 1100; compatibility decomposition for FFA1 is made in two steps, first to 3131, and then to 1100.
Same rationale for U+3132-3263.

T8. p. 369, right column, bottom

[current text]
3164 HANGUL FILLER
= cae om
≈ 1160 hangul jungseong filler
→
[proposed text]
3164 HANGUL FILLER
= chaeum
→115F hangul choseong filler
→1160 hangul jungseong filler
≈ FFA0 halfwidth hangul filler
Rationale: The usage of U3164 and UFFA0 is fairly similar. In contrast, the usage of U115F and U1160 is quite different from that of U3164 and UFFA0.

Propose partial acceptance
See disposition of comment T6.
The changes concerning compatibility mapping and cross reference is similar to the ones requested in T7 and should not be accepted for the same reason. Concerning the alias change, it could either be accepted as is or as following (no editor's preference):
= cae om, chaeum

T9. p.369, right column, bottom

[current text] Archaic letters
→
[proposed text] Old letters
Rationale: Since Hangul was invented in the 15th century, "old" seems better than "archaic". In Rep. of Korea, we use "old", not "archaic" to refer to these letters.

Propose acceptance

T10. pp. 369 ~ 370

[currently] There are annotations for code positions U3131 to U318E
→
[proposed change] Delete all annotations.
Rationale: Since the proposed changes are too drastic, we could not review carefully. Therefore we suggest to keep as in the first edition at this point. The usages of U11xx and U31xx are quite different. We will review more carefully and propose in the future.

Propose non acceptance
See disposition of comment T6.
This is a different comment but on the same repertoire already partially covered by T7 and T8 and proposes a different solution. The first edition did not show the compatibility mapping in the chart but there were nevertheless part of the standard. Removing these mapping in the new chart format would make the chart incorrect.

T11. p. 370, left column and right column

3181 ㆁ HANGUL LETTER YESIEUNG
• archaic velar nasal
≈114C ᅌ hangul choseong yesieung
3186 ㆆ HANGUL LETTER YEORINHIEUH
• archaic glottal stop
≈1159 ᅙ hangul choseong yeorinhieuh

[current text]
• archaic velar nasal
• archaic glottal stop
→
 [proposed change]
- Delete these two lines
Rationale:
1) The word archaic does not seem proper as mentioned earlier.
2) Only those two letters have more detailed information about the sound.
We could add similar information to other letters. Therefore, we propose to delete these two lines.

Propose non acceptance

Such annotations are purely editorial and have been found to be useful. Removing them because other characters have no such annotation does not seem productive. However, to be consistent with the change requested in T9, they should be changed to:
• old velar nasal
• old glottal stop

T12. p. 376, left column, top

[current text]
3200 ㈀ PARENTHESIZED HANGUL KIYEOK
≈0028 (1100 ᄀ 0029)
→
 [proposed text]
3200 ㈀ PARENTHESIZED HANGUL KIYEOK
≈0028 (3131 ㄱ 0029)
Rationale: The usages of U3131 and U1100 are quite different. For U3200, U3131 seems much better than U1100. (To represent an independent ㄱ, we need to use U1100 + U1160, not U1100 alone.)
= The same comment applies to code positions U3201 ~ U320D and U3261 ~U326D.

Propose non acceptance

See disposition of comment T6.
This would unnecessarily change compatibility mapping for all these characters, and would still result in the same compatibility decomposition because the process is recursive. The compatibility mapping content is about compatibility decomposition, not about independent representation of Jamo characters.

T13. p. 376, right column

321E PARENTHESIZED KOREAN CHARACTER O HU
≈0028 (110B ᄋ 1169 ᅩ 1112 ᄒ 116E ᅮ 0029)
[current text] O HU
→
 [proposed text] OHU
Rationale: It is a one word as in the case of U321D.

Non accepted

It is a well accepted principle that character names cannot be changed. To mitigate this, we could add in the name list either a character alias (one preceded by '※') or an annotation showing the desired character.

T14. p. 1259, right column

[current text]
FFA0 HALFWIDTH HANGUL FILLER
≈<narrow> 3164
→
 [proposed change] There seems no explanation about <narrow>, <circle>, <wide>, etc.
1) As a result, we could not review properly lines having these notations. We need to review such lines "after" explanations are given.

2) We suggest that explanations be added.
Propose acceptance
See disposition of comment T6.

T15. p. 2189, right column

[current text]
KS C 5601-1992 Korean Industrial Standards As-sociation. Jeongbo gyohwanyong buho (Code for Information Interchange).
→
[proposed text]
KS X 1001:2004 (formerly, KS C 5601), Korean Industrial Standards Association. Jeongbo gyohwanyong buhogye (Code for Information Interchange (Hangeul and Hanja)).
Propose acceptance
At the same the editor would welcome new entries for the additional Korean repertoire added after KS X 1001:2004.

T16. CJKU_SR.txt and CJKC_SR.txt

[current text]
- In CJKU_SR.txt, UTC is used
- In CJKC_SR.txt, U0- is used

→
[proposed change]
- We propose to use consistently either UTC or U0- in both CJKU_SR.txt and CJKC_SR.txt.

Propose acceptance
UTC will be used for both CJKU_SR.txt and CJKC_SR.txt. Clause 23.4 already refers to the UTCddddd format. See also resolution of comment T17.

T17. CJKC_SR.txt

[current text]
As an example, there is an entry where a UCS code position and a U0 source ref. value are the same, as shown below:
0FA0C;05140;;;;;U0-FA0C;
→
[proposed change]
- We wonder what useful information a user/reader can get from "U0-FA0C" in this example. The code positions are the same.
- There are 21 more code positions having the same situation.
- Unless U0-xxxx can give some useful information, we suggest to delete those 22 U0- entries in CJKC_SR.txt

Propose non acceptance
Code position and source reference numbers live in different coding spaces, so the fact that they have the same value is irrelevant. It probably happens for other source references. Source reference helps establish the identity of all CJK ideographs (unified or compatibility). All but one CJK ideographs have at least one source reference. The only exception is FAD4 which is in essence a deprecated character.
At the same time, it was discovered that the UTC source reference repository (UTR#45) do not contain these 22 entries and will need to be updated.

T18. 2ed CD2

- There is much change in 2ed CD and we have had hard time reviewing CD. For example, lots of new annotations, character database, etc.
- We could not finish reviewing 2ed CD and therefore plan to provide further comments in the (near?) future.
- Therefore, we suggest to make a CD2 so that member countries have enough time to review it. This way, we can make 2ed more stable and error-free.

WG2 discussion
The CD has existed in the current format since 2007 as several iterations of Working Drafts on which WG2 asked repeatedly to study and provide feedback:

- *resolution M50.37 April 2007,*
- *resolution M51.21 September 2007,*
- *resolution M52.24 April 2008,*
- *resolution M 53.27 September 2008,*
- *resolution M54.16 April 2009*

The schedule has been repeatedly pushed back, to take into account the difficulties related to CJK multicolumn production. Recognizing the size of the task at end, the SC2 secretariat and the project editor together decided to extend the CD ballot period from the required 2 months to 4 months (from 2009-05-25 to 2009-09-25).

Concerning specific points raised by the Korean NB:

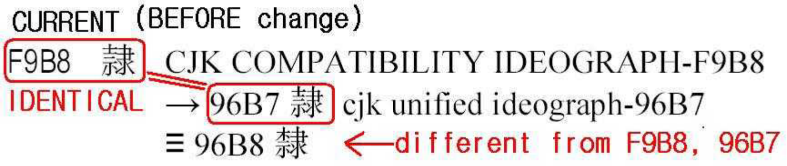
- *Annotations are informative and do not change technically the content of the standard (unless we want to make the mapping normative, but even those have been stable and are unchanged from the previous edition augmented with its amendments)*
- *The usage of the character database is targeted at simplifying the definition of common terms such as graphic characters, control characters, format character in a more sustainable way. It does not change fundamentally the technical content of the standard. The content of the database which is not used by ISO/IEC 10646 is purely informative*

Based on this, the editor does not feel that another CD is required.

T19. p. 1218, left column

(T19 and T20 are not reflecting on the CD content per say, but errata on ISO/IEC 10646)

T19.1 We request to change as shown below:

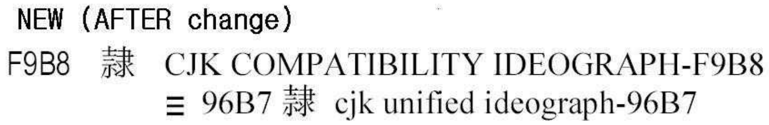


CURRENT (BEFORE change)

F9B8 隷 CJK COMPATIBILITY IDEOGRAPH-F9B8

IDENTICAL → 96B7 隷 cjk unified ideograph-96B7

≡ 96B8 隸 ←different from F9B8, 96B7



NEW (AFTER change)

F9B8 隷 CJK COMPATIBILITY IDEOGRAPH-F9B8

≡ 96B7 隷 cjk unified ideograph-96B7

<u>Propose non acceptance</u>

Canonical mapping cannot be changed in order to preserve normalization stability. There is no doubt that 96B7 would be a better mapping but it cannot be changed without destabilizing normalization which is a much larger problem. The better mapping is already shown as an annotation to F9B8 as shown above.

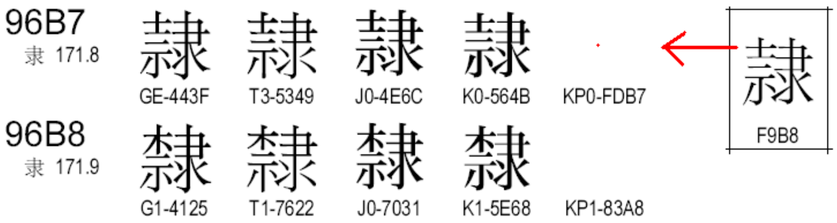
The only way out of this is to use an alternate UCS representation for the source K0-6766, either through a new compatibility CJK ideograph or through the usage of variation sequences. The rationale provided is providing more evidence of the discrepancy, but again the values for F9B8 cannot be changed.

T19.2 We request to change the following line in CJKC_SR.txt as shown below:
(current) 0F9B8;096B8;;;;K0-6766;;
→
(new) 0F9B8;096B7;;;;K0-6766;;

<u>Propose non acceptance</u>

See disposition of comment T19.1

<u>== Rationale (Information supporting our request):</u>



a) By checking the glyphs in 2ed CD, we can see that UF9B8 should be mapped to U96B7, not to U96B8.

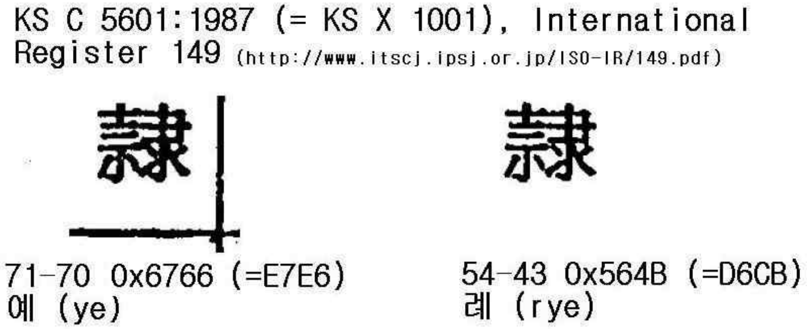
b) Furthermore, duplicate Hanja characters are included in KS X 1001 (K0), but not in KS X 1002 (K1).
- Therefore, any compatibility Hanja characters (whose source is K0, including UF9B8) must be mapped to a K0 Hanja (in this case, U96B7), but not to K1 Hanja (in this case, U96B8).

c) In CJKU_SR.txt, we know that U96B7 is a K0 Hanja and U96B8 is a K1Hanja.
096B7;GE-443F;T3-5349;J0-4E6C;K0-564B;;;KP0-FDB7;;
096B8;G1-4125;T1-7622;J0-7031;K1-5E68;;;KP1-83A8;;

d) Mapping info RE: duplicate Hanja in KS X 1001 (and comp. Hanja in UCS)
- source: Korea JTC1/SC2 documents K1645 and K1646 (= SC2/WG2 N3420 and 3421, respectively).

ro-co	KSX100	(=EUC-KR)	UCS	= ro-co	KSX1001	(=EUC-KR)	UCS
71-70	0x6766	(=E7E6)	U+F9B8	= 54-43	0x564B	(=D6CB)	U+96B7 隸

e) Exact glyphs of two Hanja characters in KS C 5601 are shown below:
- 71-70 0x6766 (=E7E6) and 54-43 0x564B (=D6CB)


KS C 5601:1987 (= KS X 1001), International Register 149 (http://www.itscj.ipsj.or.jp/ISO-IR/149.pdf)

71-70 0x6766 (=E7E6)
예 (ye)

54-43 0x564B (=D6CB)
례 (rye)

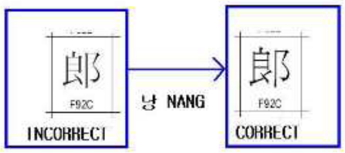
Noted

See disposition of comment T19.1

T20. pp. 1213 and 1215; CJKC_SR.txt

T20.1 p 1213: We request to change the glyph for U+F92C as shown below:

- We need to add one more stroke (i.e., The number of strokes need to be changed from 10 to 11).

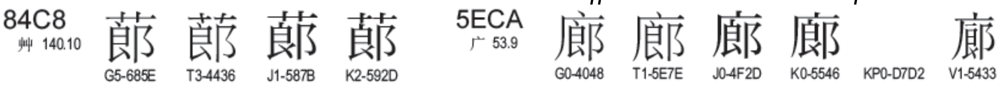


WG2 discussion

The glyph as currently shown in the chart is correctly mapped to 90CE. If you change the glyph to what NB of Korea asks, clearly it invalidates the compatibility mapping which cannot be changed for reasons given in the disposition of comment T19.1.
At the same time, the Korean NB is making a good case that the source reference K0-522B has a different glyph than the one shown for F92C (see rationale below).

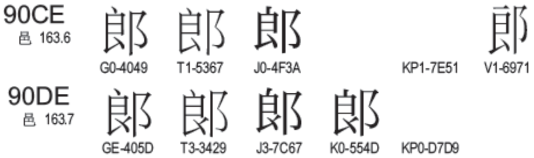
Instead of changing both the glyph and the mapping, at which point you could argue none of the current identity of the character is left, it may be better to encode another character to fix this situation. This entails the following steps:
Remove the K0-522B source from F92C, keep glyph and canonical as of now. In essence, the character becomes deprecated, in analog fashion to FAD4.
Create a new compatibility character with new glyph (11 strokes), with canonical mapping to 90DE and source reference to K0-522B.

The editor discovered that characters with similar differences have been unified elsewhere, example:



T20.2 p. 1215; We request to change two lines as shown below:
T20.2.1) Change 90CE to 90DE
<u>WG2 discussion</u>
See disposition of comment T20.1

T20.2.2) Change glyphs of two Hanja characters from 10 strokes to 11 strokes.
<u>WG2 discussion</u>
See also disposition of comment T20.1
The first character: F92C is Hanja, the second one corresponding to 90C, or 90CE, is a unified CJK Unified Ideograph with multiple sources:



Typically the glyphs corresponding to the canonical mapping have been represented using a G source font. See disposition of comment T3 from USA for detailed consideration concerning charts for CJK compatibility ideographs.

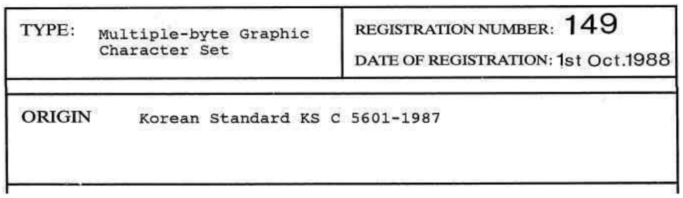
T20.3 We request to change one line in CJKC_SR.txt file as shown below:
0F92C;090CE;;;;K0-522B;;
→
0F92C;090DE;;;;K0-522B;;
<u>WG2 discussion</u>
See disposition of comment T20.1

<u>== Rationale (Information supporting out request):</u>

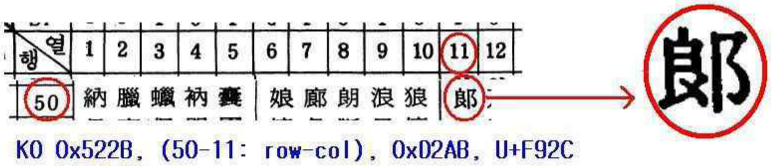
a) Mapping info RE: duplicate Hanja in KS X 1001 (and comp. Hanja in UCS)
- source: Korea JTC1/SC2 documents K1645 and K1646
(= SC2/WG2 N3420 and 3421, respectively).
낭 NANG K0 0x522B, (50-11: row-col), 0xD2AB, U+F92C
랑 RANG K0 0x554D, (53-45: row-col), 0xD5CD, U+90DE

b) Exact glyphs of two Hanja characters in KS C 5601 are shown below:
- 50-11 0x522B (=D2AB) and 53-45 0x554D (=D5CD)
- As we can see, their glyphs are exactly the same.
- source: KS C 5601-1987 (<-- International Register 149) (http://www.itscj.ipsj.or.jp/ISO-IR/149.pdf)
- The number of strokes for these two characters is 8 + 3 = 11, not 7 + 3 = 10.
(Note: The number of strokes could be 10/9 instead of 11/10. In this document, we will use 11/10).

http://www.itscj.ipsj.or.jp/ISO-IR/149.pdf

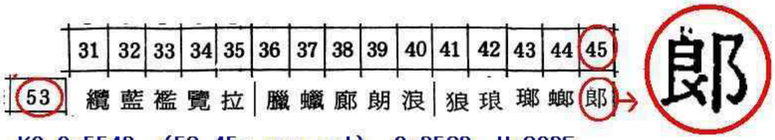


TYPE: Multiple-byte Graphic Character Set	REGISTRATION NUMBER: 149 DATE OF REGISTRATION: 1st Oct.1988
ORIGIN Korean Standard KS C 5601-1987	



행 열	1	2	3	4	5	6	7	8	9	10	11	12
50	納	臘	蠟	衲	囊	娘	廊	朗	浪	狼	郎	

K0 0x522B, (50-11: row-col), 0x02AB, U+F92C
낭 NANG



	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
53	纜	藍	襤	覽	拉	臘	蠟	廊	朗	浪	狼	琅	瑯	螂	郞

K0 0x554D, (53-45: row-col), 0x05CD, U+90DE
랑 RANG

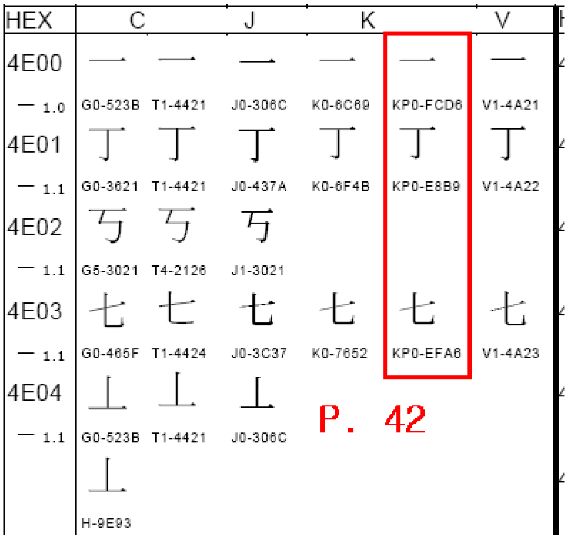
c) If the glyph of U+F92C (0x522B, 50-11, 낭 Nang) WERE correct (10 strokes),
- Since the glyph of U+F92C (0x522B) is different from the glyph of 랑 Rang (0x554D, 53-45, 11 strokes), U+F92C (0x522B) SHOULD NOT HAVE BEEN encoded as compatibility Hanja.
- Instead, we could simply fill the "currently empty" K column for U+90CE with "K0-522B/K0-5051".
- Therefore, we can conclude that the glyphs of "낭 Nang (0x522B, 50-11)" and "랑 Rang (0x554D, 53-45)" in KS C 5601 are exactly the same and the number of their strokes is 11.

d) (This is informational)
- In Ken Lunde's book, CJKV Information Processing, UF92C is correctly mapped to U90DE, which is another evidence supporting our request of change.
- However, the glyph is incorrect. We need to add one more stroke (i.e., The number of strokes need to be changed from 10 to 11).
- He promised that he would correct the glyphs.
Noted
See disposition of comment T20.1. The editor checked Ken Lunde’s books. He found the reference in page 934 of the first edition (January 1999), but not in the second edition (December 2008)

Editorial comments:

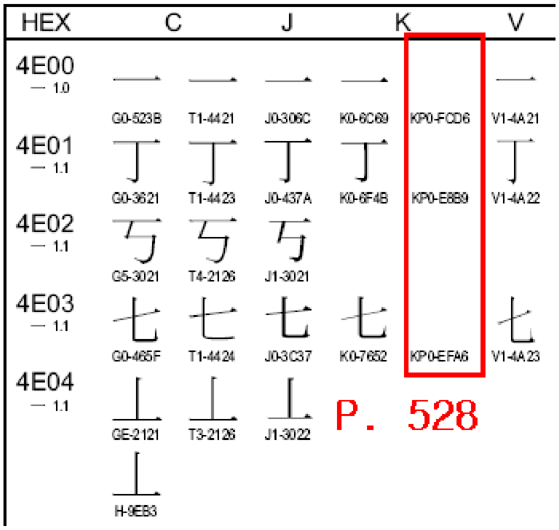
E1. pp. 42, 23.3.1 and 528.

- Due to font problems, KP columns are different on pages 42 (Fig. 2) and 528.
- We expect that fonts will solve this discrepancy.



HEX	C		J	K		V
4E00 一 1.0	一 G0-523B	一 T1-4421	一 J0-306C	一 K0-6C69	一 KP0-FCD6	一 V1-4A21
4E01 一 1.1	丁 G0-3621	丁 T1-4421	丁 J0-437A	丁 K0-6F4B	丁 KP0-E8B9	丁 V1-4A22
4E02 一 1.1	丂 G5-3021	丂 T4-2126	丂 J1-3021			
4E03 一 1.1	七 G0-465F	七 T1-4424	七 J0-3C37	七 K0-7652	七 KP0-EFA6	七 V1-4A23
4E04 一 1.1	丄 G0-523B	丄 T1-4421	丄 J0-306C			
	丄 H-9E93					

P. 42



HEX	C		J	K		V
4E00 一 1.0	一 G0-523B	一 T1-4421	一 J0-306C	一 K0-6C69	KP0-FCD6	一 V1-4A21
4E01 一 1.1	丁 G0-3621	丁 T1-4423	丁 J0-437A	丁 K0-6F4B	KP0-E8B9	丁 V1-4A22
4E02 一 1.1	丂 G5-3021	丂 T4-2126	丂 J1-3021			
4E03 一 1.1	七 G0-465F	七 T1-4424	七 J0-3C37	七 K0-7652	KP0-EFA6	七 V1-4A23
4E04 一 1.1	丄 GE-2121	丄 T3-2126	丄 J1-3022			
	丄 H-9EB3					

P. 528

<u>WG2 discussion</u>

The text in 23.3.1 is assuming that eventually a KP font will exist and is documenting it using an alternate font for the moment (note that the glyphs in page 42 are not all from NB fonts but will be updated for the next phase as more fonts become available). The alternative is to remove the KP column from the chart which should be discussed by WG2 experts.

See also disposition of comment JP29a from Japan NB.

United Kingdom: Positive with comments

Technical comments:

T.1. Clause 6.3.3 Format characters
"Code points 2060 to 206F, FFF0 to FFFC, and E0000 to E0FFF are reserved for Format Characters (see 16.3 and Annex F)."
According to the classification of code points in 6.3.1 (and also 16.3), "format" characters are characters that have a general class of Cf, Zl or Zp. Thus:
FFFC (OBJECT REPLACEMENT CHARACTER) is not a format character (gc=So)
E0100 (VARIATION SELECTOR-17) through E01EF (VARIATION SELECTOR-256) are not format characters (gc=Mn).
Proposed change:
"Code points 2060 to 206F, FFF0 to FFFB, E0000 to E00FF, and E01F0 to E0FFF are reserved for Format Characters (see 16.3 and Annex F)."
Propose acceptance in principle
The editor would even propose a more restrictive change by removing the range E01F0-E0FFF from the reserved area for format characters. The text above was derived from text in clause 8 from the previous edition. The new text would be as follows:
Code points 2060 to 206F, FFF0 to FFFB, and E0000 to E00FF are reserved for Format Characters (see 16.3 and Annex F).

T.2. Clause 16.5 Variation selectors and variation sequences
<A868, FE00> PHAGS-PA SUBJOINED LETTER reversed shaping YA
The description of the variant appearance does not match the corresponding description given in StandardizedVariants.txt:
A868 FE00; phags-pa letter reversed shaping subjoined ya
Proposed change:
Suggest changing the description to harmonize with the Unicode Standard:
<A868, FE00> PHAGS-PA LETTER reversed shaping SUBJOINED YA
Propose non acceptance
Unlike other letters in the Phags-pa list which use Phags-pa Letter, this entry uses as element a Phags-pa subjoined letter. It seems that the Unicode Standard is in error in this case.

T.3. Clause 29 Structure of the Supplementary Special-purpose Plane (SSP)
"The SSP (plane 0E) is used for special purpose use graphic characters. Code points from E0000 to E0FFF are reserved for Format Characters (see 16)."
E0100 (VARIATION SELECTOR-17) through E01EF (VARIATION SELECTOR-256) are not format characters (gc=Mn).
Proposed change:
"The SSP (plane 0E) is used for special purpose use graphic characters. Code points from E0000 to E00FF and E01F0 to E0FFF are reserved for Format Characters (see 16)."
Propose acceptance in principle
In accordance with disposition of comment T1, the new text would read as follows:
The SSP (plane 0E) is used for special purpose use graphic characters. Code points from E0000 to E00FF are reserved for Format Characters (see 16).

T.4. Annex I Ideographic description characters
"An IDS consists of an IDC followed by a fixed number of Description Components (DC). A DC may be any one of the following:
* a coded ideograph
* a coded radical
* another IDS"

Ideographic Description Sequences could usefully be applied to some non-Han scripts which comprise characters composed of one or more character components, and so it is unhelpful to restrict description components to CJK ideographs and radicals.

Proposed change:

Loosen the restrictions on what characters may be used as a Description Component in Ideographic Description Sequences, so that Ideographic Description Sequences can be used for non-CJK ideographic scripts such as Yi and other scripts such as Tangut, Jurchen and Nushu that are in the process of being standardized, as requested in WG2 N3643.

Propose acceptance in principle

Annex I is informative. The term used is 'coded ideograph', it does not say 'CJK'. To clarify the scope, we could replace in the list 'a coded ideograph' with 'a coded ideograph, CJK, Yi and others'. We cannot mention other script or block names before they are incorporated in the standard. See also comment E7 from USA.

Editorial comments:

E.1. Introduction

"This edition covers over 99 000 characters from the world's scripts."

This edition actually covers over 110,249 characters

Proposed change:

"This edition covers over 110 000 characters from the world's scripts."

Accepted

E.2. Clause 1 Scope, Note

"The Unicode Standard, Version 5.2 includes a set of characters, names, and coded representations that are identical with those in this International Standard"

This should refer to "Version 6.0".

Proposed change:

"The Unicode Standard, Version 6.0 includes a set of characters, names, and coded representations that are identical with those in this International Standard"

Accepted

E.3. Clause 2.3 Conformance of devices, Note 2

"See also 0 Annex J for receiving devices with retransmission capability"

Delete extraneous "0".

Proposed change:

"See also Annex J for receiving devices with retransmission capability"

Accepted

E.4. Clause 4.5 CC-data-element

"Unlike previous editions of the standard, this version does not use anymore implementation levels"

Suggest rewording.

Proposed change:

"Unlike previous editions of the standard, this version no longer uses implementation levels"

Accepted

E.5. Clause 6.4 Naming of characters

"The list of character names except for CJK ideographs and Hangul syllables is provided by the Unicode character Database"

"CJK ideographs" should be "CJK unified ideographs" as the Unicode character Database includes the individual names of CJK compatibility ideographs.

Proposed change:

"The list of character names except for CJK unified ideographs and Hangul syllables is provided by the Unicode character Database"

Accepted

E.6. Clause 8.1 Limited subset

"This specification allows applications and devices that were developed using other codes to inter-work with this coded character set."
"inter-work" should be spelled "interwork" (cf. 4.35 "Interworking").
Proposed change:
"This specification allows applications and devices that were developed using other codes to interwork with this coded character set."
Accepted

E.7. Clause 16.5 Variation selectors and variation sequences, Note 6

“The exhaustive list of standardized variants is also described as StandardizedVariants.html in the Unicode character database”
This note introduces a new term, "standardized variants", that is not used elsewhere in this standard. The term refers to defined variation sequences other than ideographic variation sequences, but this would not be obvious to most readers of the standard. Suggest changing the note to clarify what "standardized variants" are.
Proposed change:
“The exhaustive list of defined variation sequences other than ideographic variation sequences is also described as StandardizedVariants.html in the Unicode character database”
Accepted
In addition, the URI link should be updated from Unicode 5.0 to Unicode 6.0 at some point in the future.

E.8. Clause 23.3.2 Source reference presentatioin for CJK UNIFIED IDEOGRAPHS EXTENSION A

"The following figure shows the presentation for the CJK UNIFIED IDEOGRAPH EXTENSION A block. Up to four sources per characters are represented in a single row. If more than four sources exist, an additional row is used."
There is a maximum of three characters per row, not four.
Proposed change:
"The following figure shows the presentation for the CJK UNIFIED IDEOGRAPH EXTENSION A block. Up to three sources per characters are represented in a single row. If more than three sources exist, an additional row is used."
Accepted
See also comment T4 from Korean NB.

E.9. Annex A.1 Collection of coded graphic characters

Misplaced asterisk for:
74 UNIFIED CANADIAN ABORIGINAL SYLLABICS
116 PHONETIC EXTENSIONS SUPPLEMENT.
Proposed change:
Put the asterisk after the code point range not the collection name.
Accepted

E.10. Annex F.4 Subtending format characters

"The scope of these characters is the subsequent sequence of digits (plus certain other characters), with the exact specification as defined in the Unicode Standard, Version 5.0"
Reference to "the Unicode Standard, Version 5.0" should be updated to “Version 6.0”
Proposed change:
"The scope of these characters is the subsequent sequence of digits (plus certain other characters), with the exact specification as defined in the Unicode Standard, Version 6.0"
Accepted

E.11. Annex G Alphabetically sorted list of character names

HANGUL SYLLABLES,
CJK UNIFIED IDEOGRAPHS,
CJK UNIFIED IDEOGRAPHS EXTENSION A,
CJK UNIFIED IDEOGRAPHS EXTENSION B,

CJK UNIFIED IDEOGRAPHS EXTENSION C,
CJK COMPATIBILITY IDEOGRAPHS, and
CJK COMPATIBILITY IDEOGRAPHS SUPPLEMENT).
This omits CJK UNIFIED IDEOGRAPHS EXTENSION D
Proposed change:
HANGUL SYLLABLES,
CJK UNIFIED IDEOGRAPHS,
CJK UNIFIED IDEOGRAPHS EXTENSION A,
CJK UNIFIED IDEOGRAPHS EXTENSION B,
CJK UNIFIED IDEOGRAPHS EXTENSION C,
CJK UNIFIED IDEOGRAPHS EXTENSION D,
CJK COMPATIBILITY IDEOGRAPHS, and
CJK COMPATIBILITY IDEOGRAPHS SUPPLEMENT).
Accepted

E.12. Annex I Table I.1 Properties of ideographic description characters

Column 6 is corrupt
Proposed change:
Correct font used for column 6
Accepted
See also comment JP15. from Japanese NB.

E.13. Annex M Source of characters

Bburx Ddie Su (= Bian Xiezhe). 1984. Nuo-su bbur-ma shep jie zzit: Syp-chuo se nuo bbur-ma syt mu curx su niep sha zho ddop ma bbur-ma syt mu wo yuop hop, Bburx Ddie da Su. [Chengdu]: Syp-chuo co cux tep yy ddurx dde. Yi wen jian zi ben: Yi Han wen duizhao ban. Chengdu: Sichuan minzu chu-banshe. [An examination of the fundamentals of the Yi script. Chengdu: Sichuan National Press.]
Bburx Ddie Su. Nip huo bbur-ma ssix jie: Nip huo bbur-ma ssi jie Bburx Ddie curx Su. = Yi Han zidian. Chengdu: Sichuan minzu chubanshe, 1990. ISBN 7-5409-0128-4

These two references are incorrect. "Bburx Ddie Su" is not an author's name, but means "edited by", which was mistakenly interpreted as a proper name. The author of the 1984 work is actually given as "Syp-chuo se nuo [su] bbur-ma syt mu curx su niep sha zho ddop ma bbur-ma syt mu wo yuop hop, bburx ddie da su" which means "edited by the Sichuan Province Yi Script Working Group and Liangshan Dialect Working Committee". The author of the 1990 work is actually given as "Nip huo bbur-ma ssi jie bburx ddie curx su" which means something like "Yi-Han dictionary editorial group".
Proposed change:
Suggest following the format given in the Unicode Standard version 5.0 section R.3, where the confusing group authorship credit has been omitted:
Nuo-su bbur-ma shep jie zzit: Syp-chuo se nuo bbur-ma syt mu curx su niep sha zho ddop ma bbur-ma syt mu wo yuop hop, Bburx Ddie da Su
Nuo-su bbur-ma shep jie zzit. = Yi wen jian zi ben. Chengdu: Sichuan minzu chubanshe, 1984. Nip huo bbur-ma ssix jie. = Yi Han zidian. Chengdu: Sichuan minzu chubanshe, 1990. ISBN 7-5409-0128-4.
Accepted

E.14. Annex M Source of characters

The bibliographic references do not cover scripts added since the publication of the first edition of the standard.
Proposed change:
Suggest updating and expanding references in line with the Unicode Standard version 5.0 section R.3.
Propose acceptance in principle
The editor will look into the TUS 5.0 section R.3 and into any update available with the version 5.2. In general, proposal submitters are invited to provide relevant bibliographic references they would like to be included in annex M.

E.15. Clause 30 Code charts and list of character names

Some characters are named using American English terminology. To make the standard more useful to an international readership, add informal aliases giving British English names where appropriate.

Proposed change:

Add the British English names of the following characters as informal aliases:
2669 QUARTER NOTE
= crotchet
266A EIGHTH NOTE
= quaver
266B BEAMED EIGHTH NOTES
= beamed quavers
266C BEAMED SIXTEENTH NOTES
= beamed semiquavers
1D15D MUSICAL SYMBOL WHOLE NOTE
= semibreve
1D15E MUSICAL SYMBOL HALF NOTE
= minim
1D15F MUSICAL SYMBOL QUARTER NOTE
= crotchet
1D160 MUSICAL SYMBOL EIGHTH NOTE
= quaver
1D161 MUSICAL SYMBOL SIXTEENTH NOTE
= semiquaver
1D162 MUSICAL SYMBOL THIRTY-SECOND NOTE
= demisemiquaver
1D163 MUSICAL SYMBOL SIXTY-FOURTH NOTE
= hemidemisemiquaver, semidemisemiquaver
1D164 MUSICAL SYMBOL ONE HUNDRED TWENTYEIGHTH
= semihemidemisemiquaver, quasihemidemisemiquaver
1D13E MUSICAL SYMBOL EIGHTH REST
= quaver rest
1D13F MUSICAL SYMBOL SIXTEENTH REST
= semiquaver rest
1D140 MUSICAL SYMBOL THIRTY-SECOND REST
= demisemiquaver rest
1D141 MUSICAL SYMBOL SIXTY-FOURTH REST
= hemidemisemiquaver rest, semidemisemiquaver rest
1D142 MUSICAL SYMBOL ONE HUNDRED TWENTYEIGHTH REST
= semihemidemisemiquaver rest, quasihemidemisemiquaver rest

Propose acceptance in principle

Pending further review by the editor.

USA: Positive with comments

The U.S. National Body is voting Yes with comments on the following SC2 ballot: SC2 N4079:
CD 10646, Information technology -- Universal Coded Character Set (UCS).

Technical comments:

T.1. p 2159: "A.6 Unicode collections"

Because the reference in Annex A to Unicode collections unnecessarily duplicates information that can be derived from values specified in the Unicode character database, we suggest those collections instead be defined by reference to the Age values for Unicode characters as specified in the Unicode character database file DerivedAge.txt. Because the Age values define exactly the same lists of characters as the Unicode collections in Annex A, explicitly listing these characters in Annex A represents a maintenance burden and a potential source of error for the 10646 standard.

Propose acceptance in principle

The editor has no issue with providing this for past Unicode versions. However the file in question (DerivedAge.txt) will not be available for Unicode 6.0 until late in the process. Therefore it may be necessary to maintain a hybrid solution (DerivedAge.txt for previous versions and explicit table for future version).

T.2. CJK Unified Ideographs charts

The U.S. welcomes the publication of the whole repertoire of CJK unified ideographs in multicolumn format in the code charts, and we would like to thank the national bodies who provided fonts for that purpose. However, we are concerned that in the process of migrating to these new fonts, accidental glyph changes may have been introduced (in particular compared to the already published URO and Extension A charts). We would like WG2 to provide an assessment of the possible differences.

Propose noted

The whole point of the CD ballot is to provide an opportunity for the NBs and liaisons to submit feedback on the updated parts of the standard. The CJK repertoire is an essential part of the update. WG2 has itself no resource to do this task, unless it delegates the work to IRG members. The editor still thinks that the new chart are a vast improvement upon the previous version and does not fully understand what is requested here. A formal assessment could just add unnecessary delay to the publication of the new edition.

T.3. CJK Compatibility Ideographs charts

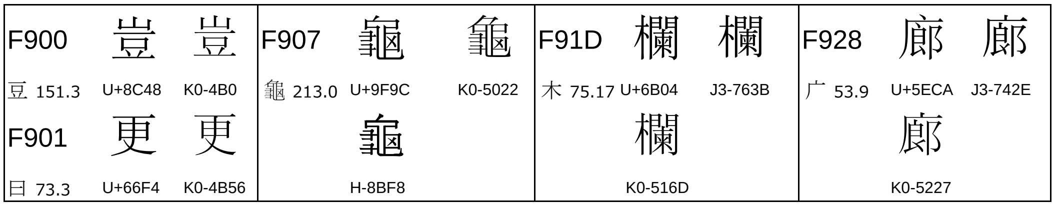
We would like the CJK compatibility ideographs code charts (both BMP and SIP) to also be presented in multi-column format. As most characters have only one or two sources, the organization adopted for Extension C would be appropriate. Furthermore, we would like the necessary fonts for those glyphs to be provided by the national bodies as well, so that they are harmonized with those of the corresponding unified ideographs.

In addition, such a change would thereby also provide multi-column charts for the 12 unified ideographs present in the CJK Compatibility Ideographs block.

Propose acceptance in principle

Using a multi-column format for CJK compatibility ideographs is in principle a good idea. However, there are some issues to consider:
- The chart currently contains a name list with explicit names specified for each character. However they are formed in a similar manner to the CJK unified ideograph and the rule could be established in a way similar to clause 24.5.
- The chart contains a canonical mapping (going to a CJK Unified Ideograph) unique to CJK compatibility ideograph. However, this could be treated from a format point of view as an additional and mandatory 'source reference', preceding the source references defined for these characters.
- There is a need to determine which glyph to be used to display the character corresponding to the canonical mapping. The editor recommends using the G source for CJK Unified Ideographs from the BMP, and the former glyph for the single column CJK Ext B when needed. There are no reference to the new extensions C and D.

- Some characters located in the CJK Compatibility Ideographs blocks are in fact CJK Unified ideographs (FA0E-FA0F, FA11, FA13-FA14, FA1F, FA21, FA23-FA24, FA27-FA29) and the information is captured through annotation in the name list. However that information is already mentioned in many other places in the standard.
- There are specific annotations for character F9B8, 2F80D, 2F814, 2F85A, 2F85B, 2F89C, 2F9B2, 2F9B6, 2F9CB, and 2F9D6. However that information is hard to spot in the name list and could be better exposed in a specific informative clause. For example, Annex I could be extended to covers character information about characters that use a chart format without an explicit name list.
- There are group headers in the name list, such as 'Pronunciation variants for KS X 1001:1998' for the range F900-FA0B, 'Duplicate characters from Big5' for the range FA0C-FA0D', etc. The information may need to be preserved, again maybe in an updated annex I.
This would allow creating a multi-column format for the CJK Compatibility Ideographs without major changes to the charting program. The figure below shows an approximation of what it could look like



F900	豈	豈	F907	龜	龜	F91D	欄	欄	F928	廊	廊
豆 151.3	U+8C48	K0-4B0	龜 213.0	U+9F9C	K0-5022	木 75.17	U+6B04	J3-763B	广 53.9	U+5ECA	J3-742E
F901	更	更		龜			欄			廊	
日 73.3	U+66F4	K0-4B56		H-8BF8			K0-516D			K0-5227	

T.4. Font contour issues in CJK multi-column charts

The font data for the glyphs for H source (3B19, 549E) and KX source (2054B, 238A7, 24C36,2597C, 270D2, 29B30) have problems with their contour orientation (in the PDFs of the code charts, rendered on screen, some of the stroke crossings are white).

Propose acceptance in principle

This is not a technical issue. The editor will request updated fonts from HKSAR and Korea (ROK) concerning these characters.

T.5. Latin Extended D

The U.S. requests that the font for Sundanese be updated in the CD, using the font as provided for the Sundanese proposal N3666 (L2/09-251).

Propose acceptance

The editor has already received the new font.

Editorial comments:

E.1. p 9, note

"The Unicode Standard, Version 5.2"

This should probably have an editor's note to remember to change "5.2" to the appropriate value. We suggest an update to "6.0."

Accepted

See also comment E2 from UK.

E.2. p 21,

"The full syntax of the notation of a short identifier, in Backus-Naur form, is { U | u } [{+}(xxxx | xxxxx | xxxxxx)]"

We recommend the removal of the "[" and "]".

Accepted

E.3. p 33, note 5:

"The variation selector only selects a different appearance of an already encoded character."

We recommend the wording be changed to: "The variation selector only selects a specific appearance among those acceptable for an encoded character."

Propose acceptance

E.4. p 34, 20.1:

“If a combining character is to be regarded as a composite sequence in its own right, it shall be coded as a composite sequence by association with the character 00AD NO-BREAK SPACE. For example, grave accent can be composed as 00AD NO-BREAK SPACE followed by 0300 COMBINING GRAVE ACCENT.”
Change 00AD to 00A0.

Propose acceptance

00AD is SOFT HYPHEN, obviously not intended. The value 00A0 corresponds to NO-BREAK SPACE.

E.5. p 34, 20.1, note:

“NOTE – Indic matras form a special category of combining characters, since the presentation can depend on more than one of the surrounding characters. Thus it might not be desirable to associate Indic matra with the character SPACE”
Change the last word from SPACE to NO-BREAK SPACE.

Propose acceptance

The note should be have been updated along with the text of the clause when SPACE was replaced with NO-BREAK SPACE in that context.

E.6. p 37, last paragraph of 22.2:

“A ‘unique-spelling’ rule is defined as follows. According to this rule, no coded character from a table for Rows 09 to 0D or 0F, or for the MYANMAR block in Row 10, shall be regarded as equivalent to a sequence of two or more other coded characters taken from the same table.
The text as it currently reads is misleading about two-part vowels. Reword the section with correct examples.

Propose acceptance

The clause 22.2 could be rewritten as follows:

Features of scripts used in India and some other South Asian countries

In the code charts for Rows 09 to 0D and 0F, and for the MYANMAR block in Row 10, of the BMP (see 30) the graphic symbols shown for some characters appear to be formed as compounds of the graphic symbols for two other characters in the same table.

EXAMPLE 1 Row 09 Devanagari

The graphic symbol for 0906 DEVANAGARI LETTER AA appears as if it is constructed from the graphic symbols for 0905 DEVANAGARI LETTER A and 093E DEVANAGARI VOWEL SIGN AA

EXAMPLE 2 Row 0D Malayalam

The graphic symbol for 0D08 MALAYALAM LETTER II appears as if it is constructed from the graphic symbols for 0D07 MALAYALAM LETTER I and 0D57 MALAYALAM AU LENGTH MARK

In such cases a single coded character may appear to the user to be equivalent to the sequence of two coded characters whose graphic symbols, when combined, are visually similar to the graphic symbol of that single character, as in a composite sequence (see 4.17).

A “unique-spelling” rule is defined as follows. According to this rule, no coded character from a table for Rows 09 to 0D or 0F, or for the MYANMAR block in Row 10, with the list of exceptions mentioned below, shall be regarded as equivalent to a sequence of two or more other coded characters taken from the same table.

- Two-part dependent vowel signs,
- Consonants including a nukta sign.

NOTE – All these characters have canonical mapping consisting of a sequence of two characters.

E.7. p. 2180 Annex I, Ideographic description characters.

The USNB requests removing the restriction on Ideographic Description Sequences that they contain only CJK ideographs and radicals to allow them to be used for other East Asian ideographic scripts.

Propose acceptance in principle

See also comment T4 from UK. Again, the editor could not find any such restriction in the current text of annex I.

E.8. New Kaithi decompositions

The U.S. requests the following Kaithi decompositions be added to the namelist for the Kaithi characters in the CD:
110AB KAITHI LETTER VA: 110A5 KAITHI LETTER BA 110BA KAITHI SIGN NUKTA
1109C KAITHI LETTER RHA: 1109B KAITHI LETTER DDHA 110BA KAITHI SIGN NUKTA
1109A KAITHI LETTER DDDHA: 11099 KAITHI LETTER DDA 110BA KAITHI SIGN NUKTA
Note: Kaithi was added in Amendment 6.

Propose acceptance