Comments on SC2 N4635, CD Text of 10646 6th Edition

Peter Constable, September 29, 2018

This document provides comments on the CD text in SC2 N4635 up through clause 5.

1. (E) — Forward

The following text reflects the fifth edition, not the sixth, and needs to be updated.

> This fifth edition of ISO/IEC 10646 cancels and replaces the fourth edition (ISO/IEC 10646:2014), which has been technically revised. It also incorporates ISO/IEC 10646:2014/Amd 1:2015 and ISO/IEC 10646:2014/Amd 2:2016.
>
> This edition includes the following significant changes with respect to the previous edition:
>
> - New scripts covered: Adlam, Bhaiksuki, , Marchen, Masaram Gondhi, Newa, Nushu, Osage, Soyombo, Tangut, and Zanabazar Square,
> - Existing scripts significantly extended: Cherokee, CJK Unified Ideographs (Extension F),
> - New Emoji symbols.

Change to:

> This sixth edition of ISO/IEC 10646 cancels and replaces the fifth edition (ISO/IEC 10646:2017), which has been technically revised. It also incorporates ISO/IEC 10646:2017/Amd 1:2018 and ISO/IEC 10646:2017/Amd 2:2019.
>
> *other text as the editor deems appropriate*

2. (E) — Clause 1, Scope

The items in the bulleted list are delimited using comma ",". However, several of the bullet items include expressions separated with commas. Normal editorial convention in such cases is for the higher-level boundaries to be separated using semi-colons. See clause 23 in ISO/IEC Directives Part 2, 8th edition, for an example that illustrates this.

Proposed change: replace commas at the end of bullet items with semi-colon.

3. (E) — Clause 1, Scope

In the bulleted list, the fourth bullet covers the BMP, and the fifth bullet covers the assigned supplementary planes. This fundamental distinction between the BMP and supplementary planes is anachronistic, a hold-over from when there were separate 10646-1 and 10646-2 standards.

Proposed change: Merge the fourth and fifth bullets into one:

- specifies the assigned planes of the UCS: the Basic Multilingual Plane (BMP), the Supplementary Multilingual Plane (SMP), the Supplementary Ideographic Plane (SIP), the Tertiary Ideographic Plane (TIP), and the Supplementary Special-purpose Plane (SSP);

(Proposed text ends with semi-colon, as proposed in comment 2.)

4 (general) — Clause 3, Terms and definitions

Links are provided to terminology databases in IEC Electropedia and ISO OBP. Terms from 10646 do not appear to be included in either of these, however.

5 (E) — Clause 3.12, code unit sequence

In note 1:

> "… any type of code points."

Change to:

> "… any type of code point."

6 (E) — Clause 3.12, code unit sequence

Note 1 refers to *types* of code points. However, this concept has not yet been introduced.

Proposed change: Add at the end of note 1, "(See clause 6.3.)"

7 (T) — Clause 3.13, collection

The definition given is:

> "numbered and named set of entities"

"Entities" is open-ended — it could include cities, unicorns, chemical formulas, etc. For the UCS context, something specific is intended, but there is no explanation of what that is. *Code points? Code point sequences?*

(Note: see related comments, below, for 3.25.)

Proposed change: Add qualifiers in the definition or in a note (new note 1) clarifying what kinds of entities are included in UCS collections. Specific, proposed wording is not provided here since it's not clear what is actually intended.

8 (E) — Clause 3.14, combining character

In note 1 (two occurrences):

> "…non-combining graphic character…"

By 3.1, "non-combining graphic character" is the same as "base character".

Proposed change: replace "non-combining graphic character" with "base character".

9 (E) — Clause 3.15, combining class

Add note:

> Note 1 to entry — See 20.2 for details on canonical ordering.

10 (E) — Clause 3.17, composite sequence

In the Unicode Standard, the corresponding term is “combining character sequence”.

Proposed change: Add “combining character sequence” as an alternate term for the same concept.

11 (E) — Clause 3.17, composite sequence

The definition includes a cross-reference, “(see also 3.14)”. Cross references should be provided in notes, not in the definition. (Per 16.5.6 of ISO/IEC Directives Part 2, “The definition shall be written in such a form that it can replace the term in its context.”)

Proposed change: Add a new note:

> “Note 1 to entry — See also 3.14.”

Renumber subsequent notes.

12 (E) — Clause 3.20, decomposition mapping

The terminology *canonical equivalent* and *compatibility equivalent* are used but are not defined in clause 3, or anywhere else in the document!

The definition of a term in clause 3 should not depend on terms that are not defined or explained within the doc. However, a thorough explanation would require details within chapter 3 of the Unicode Standard as well as UAX #15.

Proposed change: Replace the definition with the following:

> mapping from a character to a sequence of one or more characters
>
> > Note 1 to entry — Decomposition mappings are of two types: canonical decompositions, and compatibility decompositions. These are used in the derivation of various normalization forms (see 28). The code charts for various blocks include decomposition mappings and distinguish between the two types of mapping (see 33.3).

13 (E) — Clause 3.24, encoding scheme

In note 1:

> “… Some of the UCS encoding schemes have the same labels as the UCS encoding form. However, they are used in different contexts…”

Two issues:

- The definite article in “the UCS encoding form” assumes this is unique and implicitly-understood.
- The antecedent of “they” is unclear: labels for encoding schemes and encoding forms are used in different contexts? Or encoding schemes and encoding forms themselves are used in different contexts?

Proposed change:

> “… Some of the UCS encoding schemes have the same labels as UCS encoding forms. However, references to encoding schemes and encoding forms generally occur in different contexts…

14 (E) — Clause 3.25, extended collection, and note 1 of clause 3.13, collection

The definition of extended collection in 3.25 and the description of non-extended collections in 3.13 are unclear. The definition in 3.25 states (emphasis added),

> "collection for which the entities *can also* consist of sequences of code points that are in Normalization Form C"

This seems to imply that a non-extended collection must not include "sequences of code points that are in Normalization Form C". That, in turn, seems to imply that a non-extended collection *can* include sequences of code points so long as they are not in Normalization Form C (including sequences that have proper sub-sequences that are in Normalization Form C).

It's not clear if that is the actual intent, however. In 3.13, non-extended sequences are described as sets that

> "… consist only of those coded characters whose code points lie within one or more identified ranges"

That description is, itself, vague, since *any* set of coded characters code be defined using code points that "lie within one or more identified ranges", unless some constraint is imposed on "identified ranges". But it does not seem to correspond in any clear way to the definition in 3.25.

Proposed change: Give clearer definitions. Specific, proposed wording is not provided here since it is not clear what is actually intended.

15 (E) — Clause 3.28 — General Category

In note 1:

> "Each value is defined as General Category property using a two-letter abbreviation in the Unicode Standard…"

Wording is unclear. Change to:

> "Possible values are two-letter abbreviations defined for the General Category property in the Unicode Standard…"

16 (E) — Clause 3.31, high-surrogate code point

> "code point in the range D800 to DBFF reserved for the use of UTF-16"

Change to:

> "code point in the range D800 to DBFF
>
> > "Note 1 to entry — Reserved for use in UTF-16 (see 9.3)."

17 (E) — Clause 3.32, high-surrogate code unit

> "16-bit code unit in the range D800 to DBFF used in UTF-16 as the leading code unit of a surrogate pair (see 9.3)"

Change to:

> "16-bit code unit in the range D800 to DBFF and used in UTF-16

"Note 1 to entry — A high-surrogate code unit is used as the leading code unit of a surrogate pair. See also 3.40, 3.55 and 9.3."

18 (E) — Clause 3.34, ill-formed code unit sequence subset

Sets and subsets are not ordered, whereas a sequence is an ordered list of entities. The entities in a sequence can be described as coming from a set, but a sequence can include multiple instances of a given entity, whereas the set does not. Hence, the terminology "sequence subset" is odd and unclear.

This issue arises in the term being defined as well as in the definition.

Also, the definition has restrictive relative clauses introduced using "which", not "that".

Proposed changes:

- Change the term to "ill-formed code unit subsequence" (or "sub-sequence").
- Change the definition to the following.

"non-empty subsequence of a code unit sequence X that does not contain any code unit that belongs to a minimal well-formed code unit subsequence of X

"Note 1 to entry — An ill-formed code unit subsequence cannot overlap with a minimal well-formed code unit sequence."

Note: this term is only found in clause 3.61.

19 (E) — Clause 3.36, interworking

This term is not used anywhere else within the document, so it is not clear why the term is defined at all. Per ISO/IEC Directives Part 2, clause 16.5.4, only terms that are used within the document should be included in clause 3.

Proposed change: Delete this clause.

20 (E) — Clause 3.37, ISO/IEC 10646-1

"… the specification of the overall architecture and the Basic Multilingual Plane (BMP)"

Change to:

"… the specification of the overall UCS architecture and of the Basic Multilingual Plane (BMP)"

21 (E) — Clause 3.39, low-surrogate code point

"code point in the range DC00 to DFFF reserved for the use of UTF-16"

Change to:

"code point in the range DC00 to DFFF

"Note 1 to entry — Reserved for use in UTF-16 (see 9.3)."

22 (E) — Clause 3.40, low-surrogate code unit

"16-bit code unit in the range DC00 to DFFF used in UTF-16 as the trailing code unit of a surrogate pair (see 9.3)"

Change to:

> "16-bit code unit in the range DC00 to DFFF and used in UTF-16
>
> > "Note 1 to entry — A low-surrogate code unit is used as the trailing code unit of a surrogate pair. See also 3.32, 3.55 and 9.3."

23 (E) — Clause 3.44, plane

> "subdivision of the UCS codespace consisting of contiguous 65 536 code points beginning at a multiple of 65 536 which can be identified by a number from 00 to 10"

Change to:

> "subdivision of the UCS codespace consisting of 65 536 contiguous code points beginning at a multiple of 65 536
>
> > "Note to entry 1 — UCS planes can be identified by a hexadecimal number from 00 to 10"

24 (E) — Clause 3.49, row

> "subdivision of a plane consisting of contiguous 256 code points beginning at a multiple of 256 which can be identified by a number from 00 to FF"

Change to:

> "subdivision of a plane consisting of 256 contiguous code points beginning at a multiple of 256
>
> > "Note to entry 1 — Within the context of a given place, rows can be identified by a hexadecimal number from 00 to FF."

25 (E) — Clause 3.52, Supplementary Multilingual Plane for scripts and symbols

The names of planes do not include a description of the characters or blocks within the plane.

Proposed change: Change the term to "Supplementary Multilingual Plane" (i.e., remove "for scripts and symbols").

26 (E) — Clause 3.55, surrogate pair

> "representation for a single character…"

Change to:

> "UTF-16 encoded representation for a single supplementary-plane character…

27 (E) — Clause 3.59, unpaired surrogate code unit

> "code unit in a code unit sequence…"

Change to:

> "code unit in a UTF-16 code unit sequence…"

Also add a note:

"Note 1 to entry — Any unpaired surrogate code unit constitutes an ill-formed code unit sequence."

28 (E) — Clause 3.61, well-formed code unit sequence

"… and contains no ill-formed code unit sequence subset"

Change to:

"… and contains no ill-formed code unit subsequence"

See the related comment, above, on clause 3.34.

29 (T) — Clause 4.2, Conformance of information interchange

In list item a),

"… Clause 0…"

Change to "Clause 6".

30 (T) — Clause 5, General structure of the UCS

In paragraph 2,

"… from 0 to 10FFFF."

Change to:

"… from 0 to 10FFFF (hexadecimal)."

31 (E) — Clause 5, General structure of the UCS

Second item of bulleted list after paragraph 2:

"The Supplementary Multilingual Plane for scripts and symbols…"

Change to:

"The Supplementary Multilingual Plane…"

32 (T) — Clause 5, General structure of the UCS

In the paragraph after the bulleted list:

"The Tertiary Ideographic Plane (TIP, Plane 03) is reserved for ideographic characters and is currently empty."

As of this CD, the TIP is no longer empty.

Proposed changes:

- Delete this sentence from that paragraph.
- Insert a bullet item in the preceding bulleted list:

"• The Tertiary Ideographic Plane (TIP, Plane 03)."

33 (E) — Clause 5, General structure of the UCS

In the last paragraph:

> "… coding space…"

Change to:

> "… codespace…"