L2/06-316

ISO/IEC JTC1/SC2/WG2 N 3148

Date: **2006-08-09**

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

Doc. Type: **Input to ISO/IEC 10646:2003**

Title: **Proposed updates to ISO/IEC 10646:2003**

Source: **Project Editor**
Project: JTC1 02.18
Status: For review by WG2
Date: 2006-09-12
Distribution: WG2
Reference:
Medium:

This document describes updates to ISO/IEC 10646:2003 that would address issues that were discovered by the Project Editor outside the amendment process.

1. Usage of the 'composite character' term.

The sub-clause '20.4 Variation selectors' uses the term 'composite characters' without defining it. Only the term 'composite sequence' is specified in clause 4. The intent was to describe characters that are decomposable, i.e. that can be decomposed into alternate, but equivalent composite sequences. Because the term 'composite character' is only used once, it is probably not necessary to add a new formal term definition, but only to amend the sentence containing it as follows:

> Furthermore, no sequences containing variations selectors and a mix of combining characters or decomposable characters (i.e. that can be decomposed into an alternate, but equivalent composite sequence) will be defined.

It should be noted that the sentence above was moved by Amendment 3. It should also be noted that the editor instruction concerning that move is incorrect. It should say "Replace the third and fourth paragraphs with the following" instead of "Replace the *second* and *third* paragraphs…".

2. Unicode version references

In the sub-clause '20.4 Variation selectors', the 5th note contain an outdate reference to Unicode 4.0. It should be updated at minimum to Unicode 5.0, or later version, to maintain synchronization between ISO/IEC 10646 and the Unicode Standard.

In addition, the annex M should contain a new entry for Unicode 5.0 as follows:

The Unicode Consortium The Unicode Standard, Version 5.0. Reading, MA: Addison-Wesley Developer's Press, 2007. ISBN 0-321-48091-0

3. Normalization forms

The first note in clause 25 does not make much sense as written:

> NOTE 1 – By definition, the result of applying any of these normalization forms is stable over time. It means that a normalized representation of text remains normalized even when the standard is amended.

The problem is that neither the UAX #15, nor ISO/IEC 10646 explicitly implies stability. Stability is something that needs to be actively maintained by observing certain policies regarding changes to the standard. Furthermore, in the current state of affair, normalization stability is only sought for assigned characters. While new text is being drafted in the Unicode Standard to address the issue concerning normalization form stability, it seems prudent to rewrite the note mention above to reflect a better view of the situation. The new note would read as follows.

> NOTE 1 – The result of applying any of these normalization forms onto a CC-data-element is intended to stay stable over time. It means that a normalized representation of existing CC-data-element remains normalized even when the standard is amended.

This whole clause 25 (Normalization forms) will probably need to be revisited if/when UAX#15 get amended concerning stability issues.

4. Unintended Amendment 2 changes

Document SC2 N3872 which was the summary of Voting/Table of Replies on ISO/IEC 10646:2003/FPDAM2 included 'editorial' comments that were requested to be made before publication of the final text of Amendment. However some of these comments were not really editorial and changed the meaning of the affected clause. Despite the objection of the Project Editor duly transmitted to the SC2 secretariat and provided as well to ITTF, these objectionable changes were made in the final text of the amendment.

This how it was intended:

28.1.4 Name uniqueness

Each entity name must also be unique within an appropriate name space, as specified here.

Block names

Block names constitute a name space. Each block name shall be unique and distinct from all other block names specified in the standard.

Collection names

Collection names constitute a name space. Each collection name shall be unique and distinct from all other collection names specified in the standard.

Character names and named UCS sequence identifiers

Character names and named UCS sequence identifiers, taken together, constitute a name space. Each character name or named UCS sequence identifier shall be unique and distinct from all other character names or named UCS sequence identifiers.

Determining uniqueness

For block names and collection names, two names shall be considered unique and distinct if they are different even when SPACE and medial HYPHEN-MINUS characters are ignored in comparison of the names.

> NOTE 1 – A medial HYPHEN-MINUS is a HYPHEN-MINUS character that occurs immediately after a character other than SPACE and immediately before a character other than SPACE.

EXAMPLE 1 The following hypothetical block names would be unique and distinct:
LATIN-A
LATIN-B

EXAMPLE 2 The following hypothetical block names would not be unique and distinct:
LATIN-A
LATIN A
LATINA

For character names and named UCS sequence identifiers, two names shall be considered unique and distinct if they are different even when SPACE and medial HYPHEN-MINUS characters are ignored and even when the words "LETTER",

"CHARACTER", and "DIGIT" are ignored in comparison of the names.

EXAMPLE 1 The following hypothetical character names would not be unique and distinct:
MANICHAEAN CHARACTER A
MANICHAEAN LETTER A

EXAMPLE 2 The following two actual character names are unique and distinct, because they differ by a HYPHEN-MINUS that is not a medial HYPHEN-MINUS:
TIBETAN LETTER A
TIBETAN LETTER -A

The following two character names shall be considered unique and distinct:
HANGUL JUNGSEONG OE
HANGUL JUNGSEONG O-E

NOTE 2 – These two character names are explicitly handled as an exception, because they were defined in an earlier version of this International Standard before the introduction of the name uniqueness requirement. This pair is, has been, and will be the only exception to the uniqueness rule in this International Standard.

28.1.5 Annotations

A character name or a named UCS sequence identifier may be followed by an additional explanatory statement not part of the name, and separated by a single SPACE character. These statements are in parentheses and use the Latin lower case letters a-z, digits 0-9, SPACE and HYPHEN-MINUS. A capital Latin letter A-Z may be used for word initials where required.

Such parenthetical annotations are not part of the entity names themselves, and the characters used in the annotations are not subject to the name uniqueness requirements.

A character name may also be followed by a single ASTERISK separated from the name by a single SPACE. If a parenthetical annotation is present, the ASTERISK follows the annotation and is separated from the closing parenthesis by a single SPACE.

The presence of the ASTERISK notes that additional information on the character is available in annex P of this standard.

This is how it got transformed:

28.1.4 Name uniqueness

Each entity name must also be unique within an appropriate name space, as specified here.

28.1.5 Block names

Block names constitute a name space. Each block name shall be unique and distinct from all other block names specified in the standard.

28.1.6 Collection names

Collection names constitute a name space. Each collection name shall be unique and distinct from all other collection names specified in the standard.

28.1.7 Character names and named UCS sequence identifiers

Character names and named UCS sequence identifiers, taken together, constitute a name space. Each character name or named UCS sequence identifier shall be unique and distinct from all other character names or named UCS sequence identifiers.

28.1.8 Determining uniqueness

For block names and collection names, two names shall be considered unique and distinct if they are different even when SPACE and medial HYPHEN-MINUS characters are ignored in comparison of the names.

NOTE 1 – A medial HYPHEN-MINUS is a HYPHEN-MINUS character that occurs immediately after a character other than SPACE and immediately before a character other than SPACE.

EXAMPLE 1 The following hypothetical block names would be unique and distinct:
LATIN-A
LATIN-B

EXAMPLE 2 The following hypothetical block names would not be unique and distinct:
LATIN-A
LATIN A
LATINA

For character names and named UCS sequence identifiers, two names shall be considered unique and distinct if they are

different even when SPACE and medial HYPHEN-MINUS characters are ignored and even when the words "LETTER", "CHARACTER", and "DIGIT" are ignored in comparison of the names.

EXAMPLE 1 The following hypothetical character names would not be unique and distinct:
MANICHAEAN CHARACTER A
MANICHAEAN LETTER A

EXAMPLE 2 The following two actual character names are unique and distinct, because they differ by a HYPHEN-MINUS that is not a medial HYPHEN-MINUS:
TIBETAN LETTER A
TIBETAN LETTER -A

The following two character names shall be considered unique and distinct:
HANGUL JUNGSEONG OE
HANGUL JUNGSEONG O-E

NOTE 2 – These two character names are explicitly handled as an exception, because they were defined in an earlier version of this International Standard before the introduction of the name uniqueness requirement. This pair is, has been, and will be the only exception to the uniqueness rule in this International Standard.

28.1.9 Annotations

A character name or a named UCS sequence identifier may be followed by an additional explanatory statement not part of the name, and separated by a single SPACE character. These statements are in parentheses and use the Latin lower case letters a-z, digits 0-9, SPACE and HYPHEN-MINUS. A capital Latin letter A-Z may be used for word initials where required.

Such parenthetical annotations are not part of the entity names themselves, and the characters used in the annotations are not subject to the name uniqueness requirements.

A character name may also be followed by a single ASTERISK separated from the name by a single SPACE. If a parenthetical annotation is present, the ASTERISK follows the annotation and is separated from the closing parenthesis by a single SPACE.

The presence of the ASTERISK notes that additional information on the character is available in annex P of this standard.

In effect, the previous content of sub-clause 28.1.4 got pulled out and made into new sub-clauses. This was an original misunderstanding from the commenter and resulted in a significant technical change of the standard by the ITTF editor. The request is to put it back as it was intended and technically approved by the National Bodies. To avoid further confusion, the following titles: Block names, Collection names, Character names and named UCS sequence identifiers, and Determining uniqueness can be formatted differently (not bold, underlined).

5. CJK source information

Document SC2 N3814 which was the summary of Voting/Table of Replies on ISO/IEC 10646:2003/FPDAM1 included comments concerning CJK Ideograph source information. Some were clear errata that could be acted upon immediately, others were clarification requests which required more times. The clear errata concerned Hong Kong Supplementary Character Set source references and the appropriate correction were made in Amendment 1 before publication. After review by WG2/IRG and production of the document WG2 N3132 (IRG N1218) it is now possible to complete the remaining corrections. There are as follows:

a) The source reference info for U+4695 鶯 is changed from:

```
04695;G3-4862;T4-6E3B;;K3-322D;;;KP1-752A;
```

to:

```
04695;G3-4862;;;K3-322D;;;KP1-752A;
```

This means that U+278AE 鶯 remains the only character having T4-6E3B as source reference as in:

```
278AE;G_HZ;T4-6E3B;;;;;;
```

b) The source reference info for U+FA23 赴 is changed from:

0FA23;;TF-3862;JA-2728;;;;;U0-FA23

to:

0FA23;;;JA-2728;;;;;U0-FA23

This means that U+27EAF 𧺯 remains the only character having TF-3862 as source reference as in:

27EAF;;TF-3862;;;;;;

(U+27EAF is really a unification error, but is already referenced in mapping to TF-3862 in CNS 11643)

c) The source reference info for U+4443 朦 is changed from:

04443;G_KX;T3-5866;;K3-3039;;;KP1-6C04;

to:

04443;G_KX;T3-5866;;K3-3039;V0-417A;;KP1-6C04;

This means that U+6726 朦 remains the only character having V0-417A as source reference as in:

06726;G0-6B7C;T1-764D;J0-5B2F;K0-5954;V0-417A;;KP0-DBC2;

d) The source reference info for U+6F58 潘 is changed from:

06F58;G0-454B;T1-6D59;J0-5F2F;K0-5A6B;V2-8D4D;;KP0-DCF3;

to:

06F58;G0-454B;T1-6D59;J0-5F2F;K0-5A6B;;;KP0-DCF3;

This means that U+6DFB 添 remains the only character having V2-8D4D as source reference as in:

06DFB;G0-4C6D;T1-5B50;J0-453A;K0-7455;V2-8D4D;;KP0-EDC7;

e) The source reference info for U+24319 𤌍 is changed from:

24319;G_FZ_BK;;;;;;;

to:

24319;G_FZ;;;;;;;

The original intent was to show a double source (G_FZ and G_BK), because this format is not documented and based on recommendation from IRG N1218, the source is modified into ‘G_FZ’.

These changes, if adopted for amendment 3, would require re-publishing the file CJKU_SR.txt which contains the CJK Unified Ideograph source information.

6. Mirroring characters

Annex E contains all mirrored characters in bidirectional context. In addition, according to the note included in clause 19, the list should represent all characters which have the ‘Bidi Mirrored’ property in the Unicode Standard. Comparing these two sets of characters, using the property value mentioned above as defined in Unicode 5.0, provides the following result:

a) Characters not in ISO/IEC 10646:2003 + Amendment 1 and 2 but in Unicode 5.0

```
0F3A   TIBETAN MARK GUG RTAGS GYON
0F3B   TIBETAN MARK GUG RTAGS GYAS
0F3C   TIBETAN MARK ANG KHANG GYON
0F3D   TIBETAN MARK ANG KHANG GYAS
169B   OGHAM FEATHER MARK
169C   OGHAM REVERSED FEATHER MARK
2018   LEFT SINGLE QUOTATION MARK
2019   RIGHT SINGLE QUOTATION MARK
201A   SINGLE LOW-9 QUOTATION MARK
```

```
201B   SINGLE HIGH-REVERSED-9 QUOTATION MARK
201C   LEFT DOUBLE QUOTATION MARK
201D   RIGHT DOUBLE QUOTATION MARK
201E   DOUBLE LOW-9 QUOTATION MARK
201F   DOUBLE HIGH-REVERSED-9 QUOTATION MARK
301D   REVERSED DOUBLE PRIME QUOTATION MARK
301E   DOUBLE PRIME QUOTATION MARK
301F   LOW DOUBLE PRIME QUOTATION MARK
FE59   SMALL LEFT PARENTHESIS
FE5A   SMALL RIGHT PARENTHESIS
FE5B   SMALL LEFT CURLY BRACKET
FE5C   SMALL RIGHT CURLY BRACKET
FE5D   SMALL LEFT TORTOISE SHELL BRACKET
FE5E   SMALL RIGHT TORTOISE SHELL BRACKET
FE64   SMALL LESS-THAN SIGN
FE65   SMALL GREATER-THAN SIGN
1D6DB  MATHEMATICAL BOLD PARTIAL DIFFERENTIAL
1D715  MATHEMATICAL ITALIC PARTIAL DIFFERENTIAL
1D74F  MATHEMATICAL BOLD ITALIC PARTIAL DIFFERENTIAL
1D789  MATHEMATICAL SANS-SERIF BOLD PARTIAL DIFFERENTIAL
1D7C3  MATHEMATICAL SANS-SERIF BOLD ITALIC PARTIAL DIFFERENTIAL
```

b) Characters not in Unicode 5.0 but in ISO/IEC 10646:2003 + Amendment 1 and 2

```
27C8   REVERSE SOLIDUS PRECEDING SUBSET
27C9   SUPERSET PRECEDING SOLIDUS
```

To preserve consistency between the two standards, the characters in list a) should be added to Annex E and the characters in list b) should be removed from Annex E. Note that the comparison was done using automated tools (Perl based) to avoid human errors.

7. Removal of levels from the standard

Clearly, the implementation level concept used in ISO/IEC is becoming rapidly obsolete for the following reasons:

- No major implementation is using anything but level 3 which is the unrestricted level.
- The Unicode Standard only references implementation level 3.
- Every new character submitter is baffled by the level question asked in the Proposal Summary Form, especially when the proposed repertoire contains combining characters. Should they use level 2 or level 3? Only a detailed reading of the standard itself can really answer that question.
- Combining marks have not been added to the level 2 restricted repertoires for many years.
- There is a new Proposal Summary Form being evaluated that would eliminate altogether the level question from the form.

The proposal is to eliminate the choice of multiple implementation levels from new versions of the standard and limit implementation level to level 3. Implementations requiring the former implementation levels 1 and 2 could still refer to the standard up to ISO/IEC 10646:2003 and the two current amendments. For identification purpose, the sole remaining level: level 3 would still be mentioned in places such as ISO/IEC 2022, to maintain compatibility with current versions of ISO/IEC 10646. The following is a fairly comprehensive set of editor's instructions to achieve that goal:

Page 1, sub-clause 2.2 Conformance of information interchange

In second paragraph, remove ', and to an identified implementation level chosen from clause 14'.

In fifth paragraph, remove ', the adopted implementation level'.

Page 1, sub-clause 2.3 Conformance of devices

In second paragraph (after the note), remove 'the adopted implementation level,'.

In fourth and fifth paragraph (b and c statements), remove ‘and implementation level’.

Page 11, clause 14 Implementation levels

Replace content as following:

Unlike previous editions of the standard, this version does not use anymore various implementation levels. It only uses one level which was formally referenced as ‘Implementation level 3’ and as such may contain coded representations of any characters. To maintain compatibility with these previous editions, in the context of identification of coded representation in standards such as ISO/IEC 2022, the implementation level may still be referenced as ‘Implementation level 3’. See clause 16.2.

Page 12, sub-clause 16.1 Purpose and context of identification

In *first paragraph, remove* ‘, the implementation level,’.

In second paragraph, remove ‘with an implementation level’.

Page 12, sub-clause 16.2 Identification of UCS coded representation form with implementation level

Rename sub-clause ‘Identification of UCS coded representation form’.

In first paragraph, remove ‘and an implementation level (see clause 14)’.

Replace the 6 item list by the following 2 item list:

ESC 02/05 02/15 04/05
UCS-2 with implementation level 3
ESC 02/05 02/15 04/06
UCS-4 with implementation level 3

Page 18, clause 24 Combining characters

Replace first paragraph with the following:

This clause specifies the use of combining characters. A list of combining characters is shown in clause B.

Page 18, sub-clause 24.3 Alternate coded representations

In note, remove ‘in implementation level 3’.

Page 19, sub-clause 24.5 Collections containing combining characters

Remove second paragraph, starting with ‘When implementation level’.

In third (last) paragraph, remove last sentence, starting with ‘Such a collection’.

Page 20, sub-clause 26.1 Hangul syllable composition method

Remove third paragraph, starting with ‘The implementation level’.

Page 20, sub-clause 26.2 Features of scripts used in India and some other South Asian countries

Remove last paragraph, starting with ‘This “unique-spelling” rule shall apply’ *and following note*.

Page 1352, annex A.1 Collection of coded graphic characters

In second note, remove first sentence, starting with ‘Use of implementation level’.

Page 1360, annex B List of combining characters

Remove header B.1 List of all combining characters

Remove sub-clause B.2 List of combining and other characters not allowed in implementation level 2.

Page 1368, annex C.5 Identification of UTF-16

Replace first paragraph and list by the following:

When the escape sequences from ISO/IEC 2022 are used, the identification of UTF-16 shall be by the following designation sequence:

ESC 02/05 02/15 04/12
UTF-16 with implementation level 3

Page 1373, annex D.6 Identification of UTF-8

Replace first paragraph and list by the following:

When the escape sequences from ISO/IEC 2022 are used, the identification of UTF-8 shall be by following designation sequence:

ESC 02/05 02/15 04/09
UTF-8 with implementation level 3

Page 1395, annex N.1 Methods of reference to character repertoires and their coding

In second paragraph, including the list, replace 'the adopted implementation level (1,2 or 3)' *with* 'the implementation level 3 (see clause 14)'

Page 1395, annex N.2 Identification of ASN.1 character abstract syntaxes

Replace text after the first note with the following:

The first such arc following a 10646 arc identifies the implementation level, and is level-3 (3).

NOTE 2 – This version of the standard only specifies the implementation level 3. See clause 14.

The second such arc identifies the repertoire subset, and is either:
- all (0), or
- collections (1).

Arc (0) identifies the entire collection of characters specified in ISO/IEC 10646. No further arc follows this arc.

NOTE 3 – This collection includes private groups and planes, and is therefore not fully-defined. Its use without additional prior agreement is deprecated.

Arc (1) is followed by one or a sequence of further arcs, each of which is a collection number from annex A, in ascending numerical order. This sequence identifies the subset consisting of the collections whose numbers appear in the sequence.

NOTE 4 – As an example, the object identifier for the subset comprising the collections BASIC LATIN, LATIN-1 SUPPLEMENT, and MATHEMATICAL OPERATORS, is:

{iso standard 10646 0 level-3 (1) collections (1) 1 2 39}

ISO/IEC 8824 also specifies object descriptors corresponding to object identifier values. For each combination of arcs the corresponding object descriptors are as follows:

3 0 : "ISO 10646 level-3 unrestricted"

For a single collection with collection name "xxx".

3 1 : "ISO 10646 level-3 xxx"

For a repertoire comprising more than one collection, numbered m1, m2, etc.

3 1 : "ISO 10646 level-3 collections m1, m2, m3, .. "

NOTE 5 – All spaces are single spaces.
