WG2 N5174

Date: 2022-01-31

Title: Proposed changes concerning Character Name Aliases in ISO/IEC 10646

Source: Michel Suignard, Project editor
Status: Individual Contribution
Distribution: WG2
Reference: N5168

Executive Summary:

This document presents proposed changes to ISO/IEC 10646 to address concerned raised in document N5168 (https://www.unicode.org/wg2/docs/n5168R1-ISO10646.pdf) . The main modification is to make the file NameAliases.txt which is currently part of every version of the Unicode Standard a normative part of ISO/IEC 10646 by reference. It will require some changes in various part of the standard which are detailed below. In addition, some minor changes will be done to address UTF-16 byte order considerations also made in N5168.

Character Name Aliases

Initially character names aliases were introduced in ISO/IEC 10646 to clarify the use of the '※' notation in the code chart which is described in sub-clause 34.3 'Character names list' as a marker for the normative character name alias. The concept is itself introduced in sub-clause 7.4 'Naming of characters' and is again incorporated in the names specification in sub-clause 27.1 'Entity name'. Finally, the related data file: NamesAliases.txt is mentioned in a Note in clause 8 'Revision and updating of the UCS'.

As currently noted by the authors of N5168 'Names aliases and UTF-16 encoding scheme ..', the current situation creates some divergence between ISO/IEC 10646 and the Unicode Standard by not fully capturing the various uses of the data file, some being pertinent to ISO/IEC 10646, such as formal names related to the C0/C1 control function.

The recommendation indicated by the N5168 authors to fully synchronize the name aliases from the UCD NameAliases.txt into ISO/IEC 10646 can be implemented by doing the changes documented below.

Control character 'names'

The authors of N5168 claim that 'in the Unicode Standard, names for control character are provided via NamesAliases in the UCD.' Furthermore, it also says that 'However, in ISO 10646, control characters have no normative names.'. However, this is not an exact representation of the current status. In Unicode, along with Private-Use, Surrogate, Non-character, and Reserved, Control characters have a null string as value for their Name property (reference NR4 in section 4.8 'Name' in the Unicode v14.0 Core specification). In Unicode, the Name property is what defines the character name associated with a code point. In Unicode, C0/C1 code points have no string associated with their Name property, unlike other typical assigned code points.

The mechanism in NameAliases provides various types of aliases, including a 'control' type associating code points in the ranges covered by C0/C1 with ISO6429 names for these control functions. These aliases provide immutability and uniqueness in the overall name space used in the combined set of Unicode name covered by the Name property and the character name aliases covered by the NameAliases.

The rationale concerning the relationship between C0/C1 code points and formal names is beyond the scope of this document and has been debated for a long time. What can be achieved at this point is to synchronize the status between ISO/IEC 10646 and the Unicode standard.

Proposed change to ISO/IEC 10646 for character name aliases

1) Make the file NameAliases.txt a normative part of ISO/IEC 10646 by adding it to clause 2 'Normative references': (Unicode version to be updated to the version associated with the future amendment, likely version 15.0)

   Unicode Standard Version 14.0, Character Name Aliases:
   https://www.unicode.org/Public/14.0.0/ucd/NameAliases.txt

2) Modify the content of sub-clause 7.4, paragraph after item g) as follows:
   'Some characters may have one or more alternate names, called character name aliases. See 7.5.'
   The following note is preserved:

   NOTE — Character name aliases, which are normative, should not be confused with informative aliases, which are other names for characters that may be used outside this document but that are not normative

3) Introduce a new sub-clause 7.5 Character name aliases:

   7.5 Character name aliases
   This document has a mechanism for the publication of additional, normative formal aliases for characters. These formal aliases are known as character name aliases. They function essentially as auxiliary names for a character. Their main usage is to provide correction for known mistakes in character names, but they have other usages such as providing string identifiers for control functions.
   Character name aliases are listed in the file NameAliases.txt (see Clause 2). That file also documents the 'type' field which distinguishes among different kinds of character name aliases, as shown in Table 2.

   Table 2: Types of Character Name Aliases

Type	Description
correction	Corrections for mistakes in the character names which cannot be fixed after publication of the standard
control	ISO/IEC 6429 names for C0 and C1 control functions, and other commonly occurring names for these control functions
alternate	Widely used alternate names for format characters
figment	Several documented labels for C1 control code points which were never actually approved in any standard
abbreviation	Commonly occurring abbreviations or acronyms for control functions, format characters, spaces and variation selectors

   Character name aliases follow the same rules as character names. See Clause 27.

   (Other tables and following sub-clauses renumbered as appropriate)

4) Remove the Note in clause 8 (now superfluous):

   NOTE – Character name aliases are created to denote errors in the character names which cannot be fixed after publication of the standard. These character name aliases are described in the file NameAliases.txt part of the Unicode character database (http://www.unicode.org/Public/UCD/latest/ucd/NameAliases.txt).

5) In clause 12 'Use of control function with the UCS' replace the current NOTE 3:

NOTE 3 – The following list provides the long names from ISO/IEC 6429 used in association with the control characters.

0000 NULL
0001 START OF HEADING
0002 START OF TEXT
0003 END OF TEXT
0004 END OF TRANSMISSION
0005 ENQUIRY
0006 ACKNOWLEDGE
0007 BELL
0008 BACKSPACE
0009 CHARACTER TABULATION
000A LINE FEED
000B LINE TABULATION
000C FORM FEED
000D CARRIAGE RETURN
000E SHIFT-OUT
000F SHIFT-IN
0010 DATA LINK ESCAPE
0011 DEVICE CONTROL ONE
0012 DEVICE CONTROL TWO
0013 DEVICE CONTROL THREE
0014 DEVICE CONTROL FOUR
0015 NEGATIVE ACKNOWLEDGE
0016 SYNCHRONOUS IDLE
0017 END OF TRANSMISSION BLOCK
0018 CANCEL
0019 END OF MEDIUM
001A SUBSTITUTE
001B ESCAPE
001C INFORMATION SEPARATOR FOUR
001D INFORMATION SEPARATOR THREE
001E INFORMATION SEPARATOR TWO
001F INFORMATION SEPARATOR ONE
007F DELETE
0082 BREAK PERMITTED HERE
0083 NO BREAK HERE
0084 INDEX
0085 NEXT LINE
0086 START OF SELECTED AREA
0087 END OF SELECTED AREA
0088 CHARACTER TABULATION SET
0089 CHARACTER TABULATION WITH JUSTIFICATION
008A LINE TABULATION SET
008B PARTIAL LINE FORWARD
008C PARTIAL LINE BACKWARD
008D REVERSE LINE FEED
008E SINGLE-SHIFT TWO
008F SINGLE-SHIFT THREE
0090 DEVICE CONTROL STRING
0091 PRIVATE USE ONE
0092 PRIVATE USE TWO
0093 SET TRANSMIT STATE
0094 CANCEL CHARACTER
0095 MESSAGE WAITING
0096 START OF GUARDED AREA
0097 END OF GUARDED AREA
0098 START OF STRING
009A SINGLE CHARACTER INTRODUCER
009B CONTROL SEQUENCE INTRODUCER
009C STRING TERMINATOR
009D OPERATING SYSTEM COMMAND
009E PRIVACY MESSAGE
009F APPLICATION PROGRAM COMMAND

The control character 0084 INDEX has been removed from ISO/IEC 6429. In addition, the control characters 000E and 000F are named SHIFT-OUT and SHIFT-IN respectively in 7-bit environment and LOCKING-SHIFT ONE and LOCKING-SHIFT ZERO respectively in 8-bit environment.

by the following text and note:

Control functions have associated string identifiers specified by the control type in the file NamesAliases.txt (see Clause 2). Many of these control functions have multiple identifiers based on various environments.

NOTE 3: For example, the control characters 000E and 000F are named SHIFT-OUT and SHIFT-IN respectively in a 7-bit environment and LOCKING-SHIFT ONE and LOCKING-SHIFT ZERO respectively in an 8-bit environment.

Proposed change to ISO/IEC 10646 for UTF-16 byte order

The document is asking for a minor change to the sub-clause 11.5 ‘UTF-16’ to keep the Unicode Standard and ISO/IEC 10646 synchronized. This seem not controversial and should be part of the next amendment text. Therefore the 3rd paragraph would read as follows (italicized text added):

In the absence of signature *or a higher-level protocol*, the octet order of the UTF-16 encoding scheme is that the more significant octet precedes the less significant octet.

-end of document-