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

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

Title: Improving formal definition for control characters

Source: Project Editor
Project: JTC1 02.18
Status: For review by WG2
Date: 2006-02-22
Distribution: WG2
Reference:
Medium:

Summary: The definition for control characters in ISO/IEC 10646 is lacking. They are not defined in the term sections, have no formal names and are only normatively introduced by referencing ISO 6429. The situation is creating issues for organization such as ITU.

The control characters in ISO/IEC 10646 (0000-001F, 007F, 0080-009F) have many oddities in the standard:
-   they are not formally defined in the term definition section,
-   they do not have formal names (they only acquire them indirectly by reference to ISO/IEC 6429),
-   they do not belong to any blocks,
-   they appear randomly in collections, for example, they belong to collection 300 (BMP), but not to 302 (BMP SECOND EDITION, 340 (COMBINED FIRST EDITION),
-   some have been formally deprecated by ISO/IEC 6429 (DEL, IND),
-   some are not defined (0080, 0081, 0099),
-   some have two names/meaning (000E and 000F) depending on their usage on a 7 or 8 bit environment.

The following list shows all these characters based on ISO/IEC 6429 (aka ECMA-48 at http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-048.pdf):

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 DEL
0080 CONTROL CODE C1-80 *
0081 CONTROL CODE C1-81 *
0082 BREAK PERMITTED HERE

0083 NO BREAK HERE
0084 IND
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
0099 CONTROL CODE C1-99 *
009A SINGLE CHARACTER INTRODUCER
009B CONTROL SEQUENCE INTRODUCER
009C STRING TERMINATOR
009D OPERATING SYSTEM COMMAND
009E PRIVACY MESSAGE
009F APPLICATION PROGRAM COMMAND

* Note that the names for 0080, 0081, and 0099 are new. Possibly 0084 should use a similar naming scheme because it is deprecated by ISO/IEC 6429, although it is commonly known as 'IND', so this seems a better approach.

It seems that they should be formally defined in clause 4 (terms and definitions), with the following text (derived from ISO/IEC 6429):

Control character (new)
A control function the coded representation of which consists of a single code position.

Control function (modified)
An action that affects the recording, processing, transmission, or interpretation of data, and that is represented by a CC-data-element.

Unique names should be created for the code positions that don't have any. A block and a collection should be created and they should be added to collections corresponding to future editions of the standard.

The second paragraph of clause 8 which describes the code positions as reserved for the control characters can be simply removed.

This would simply add names to these control characters, but would not add any new functionality beyond what is already described in 10646. Other standards using control characters are not necessarily affected by this revision; they can simply refer to ISO/IEC 6429. However, it would make it easier for standard bodies like ITU to reference 10646 when they need to simply mention that they are using 10646, including the control characters that are located in the C0 and C1 area. Finally, should the control character behavior needs further description in 10646, having formal names would make such addition easy to process.