

MES-1 (ISO 10646 character collection 281) considered defective

Karl Pentzlin - 2008-08-05

The following is an excerpt from:

*DRAFT - An alternative to the current ISO/IEC 9995-3
Draft Version 9a – 2008-07-31
DIN NA 043-01-35-01 GAK, document N604*

mentioned in UTC document L2/08-278 as being found on:

<http://www.csc.fi/english/pages/meek/An-alternative-to-the-current-ISO-IEC-9995-3>

As this link may be temporary, the mentioned part is presented as a separate document here.

The excerpt references to:

- MES-1 (Multilingual European Subset 1):
collection 281 (titled MES-1) as specified in amendment 1 to ISO/IEC 10646:1-2000

A6. Notes on the previous version of ISO/IEC 9995-3

The current version of ISO/IEC 9995-3 intends to enable the input of a character repertoire as defined by collection 281 (MES-1) specified in amendment 1 to ISO/IEC 10646-1:2000.

In the current version of Part 2 of ISO/IEC 9995 is stated:

For the input of graphic character repertoire of collection 281 (titled MES-1) as specified in amendment 1 to ISO/IEC 10646:1-2000, a Common Secondary Group Layout (to be used as group 2) is specified in ISO/IEC 9995-3.

The collection 281 is:

U+00..: 20-7E A0-FF
U+01..: 00-13 16-2B 2E-4D 50-7E
U+02..: C7 D8-DB DD
U+20..: 15 18-19 1C-1D AC
U+21..: 22 26 5B-5E 90-93
U+26..: 6A

Today, this character collection is not suited as base for a character repertoire for a small international keyboard extension (by only a single group), for the following reasons:

- The collection 281 is based of the ISO/IEC 6937, which was developed in the 1970s for "telematic services", i.e. for communication purposes like the long outdated Telex successor "Teletex". It was not its primary goal to act as a well thought set for an international keyboard.
- ISO/IEC 6937 does not adhere to the same encoding principles as Unicode, which is prevalent in today's data processing systems. Especially, it is missing the mechanism of Unicode's combining characters. The ISO/IEC 6937 conforming mechanism of forming an accented letter by base letter + backspace + spacing accent is no longer possible with Unicode.
- Moreover, some characters of the collection 281 are obsolete legacy today which do not need to burden an actual keyboard design.
- The last 30 years yielded the need for some more characters (e.g., the introduction of the Latin alphabet in Azerbaijan revived the Janjalif character Ə/ə).
- Additionally, the collection 281 is defective (e.g., it contains the characters Đ/đ, Ʀ/Ƨ and Ɖ/ɖ for Sami, but not Ǫ/ǫ, Ǟ/ǟ and Ǡ/ǡ).
- As the name "MES-1" ("Multilingual European Subset 1") suggests, the larger part of the world is not considered (especially Vietnamese, but also most "minority languages" even if they write Latin).
— *When Sami is taken into account in an international standard, why not Yorùbá or Comanche?*

Especially, the following characters appear to be of inferior priority, superfluous, or even wrong:

U+00A6 BROKEN BAR
No real use attested beyond special mathematical-logical applications. Historic variant of U+007C
VERTICAL LINE.

U+00AC NOT SIGN
Mathematical symbol without any attested business use. No need for inclusion into a repertoire which does not cover an appropriate large set of symbols for mathematics or formal logic.

U+0132 LATIN CAPITAL LIGATURE IJ
U+0133 LATIIN SMALL LIGATURE IJ
These are nowadays written as separate letters (see the detailed discussions in the Unicode mailing list archive).

U+0138 LATIN SMALL LETTER KRA
Was used in a former Greenlandic orthography, now obsolete there.

U+013F LATIN CAPITAL LETTER L WITH MIDDLE DOT
U+0140 LATIN SMALL LETTER L WITH MIDDLE DOT
Included in Unicode only as legacy compatibility characters. The preferred representations for Catalan by Unicode are U+004C U+00B7 resp. U+006C U+00B7.

U+0149 LATIN SMALL LETTER N PRECEDED BY APOSTROPHE
Included in Unicode only as legacy compatibility character.
The appropriate Unicode representation of the Afrikaans letter 'n is U+02BC U+006E.

U+2126 OHM SIGN
This Unicode character has a canonical equivalence to U+03A9 GREEK CAPITAL LETTER OMEGA. Therefore according to the Unicode rules, the latter code is to be preferred for the Ohm sign.
(Note: To the U+00B5 MICRO SIGN, such an argument does not apply, as it has only a compatibility equivalence to U+03BC GREEK SMALL LETTER MU.)

U+266A EIGHTH NOTE
No special use is attested for this symbol in plain text. No need for inclusion into a repertoire which does not cover an appropriate large set of iconic symbols.

U+00...: C0-C5 C7-CF D1-D6 D9-DD E0-E5 E7-EF F1-F6 F9-FD FF
U+01...: 00-0F 12-13 16-25 28-2B 2E-2F 34-37 39-3E 43-4D 50-51 54-65 68-7E
These are precomposed letters, not needing to be enumerated as it is sufficient to have the constituent characters to be included in the repertoire
All precomposed letters are implicitly contained in a set which includes all characters generated by application of the Unicode Normalization Form NFC on any sequence of characters specified in the original repertoire.