Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: Working Group Document
Title: Proposal to encode missing Latin small capital and modifier letters in the UCS
Source: ISO/IEC JTC1/SC35
Authors: ISO/IEC JTC1/SC35/WG1
Status: Liaison Contribution
Action: For consideration by JTC1/SC2/WG2 and UTC
Date: 2011-03-12

1. Introduction

1.1 The upcoming standard "Multilingual, Multiscript Keyboard Group Layouts"

ISO/IEC JTC1/SC35/WG1 currently is working on a new standard ISO/IEC 9995-9 "Multilingual, Multiscript Keyboard Group Layouts".

This standard describes methods to input a large set of characters by keyboards which employ 26 keys labeled by the basic Latin alphabet A...Z, besides a minimal set of other required keys.
While this requirement is fulfilled by several PC keyboards (whether they are designed primarily for a Latin written language, or employ the Latin letters by a secondary group on layouts designed primarily for a language written in another script), ISO/IEC 9995-9 explicitly also aims to the "miniature keyboards" which are commonly found on mobile phones and similar small mobile devices (whether supplied physically or presented on-screen).

The basic mechanism is "switching" (permanently) or "latching" (affecting only the next key actuation) to another "group" of 52 characters, which are mapped to the capital and small basic Latin letters A...Z + a...z.
Subsequently, the characters of that "group" are then input by pressing the keys labeled by that Latin letters in the shifted or unshifted state. (This "switching" or "latching" is done by actuating a "superselect" appliance, which may e.g. be a key combination like "AltGr + Tab" on a PC keyboard, or "Shift" + Symbol key" on a miniature keyboard, followed by pressing one or more Latin labeled keys to select the wanted "group").

The goal of the standard is to enable the input of all characters needed for a specific set of languages using the same script, as well of all characters which occur in plain text within a specific field of application, in a user-friendly way.

1.2 The role of modifier letters within the upcoming standard ISO/IEC 9995-9

Small capital and superscript letters are part of some orthographies. In addition, subscript small letters are plain text characters in certain fields of application (especially linguistics). Therefore, they are in the scope of ISO/IEC 9995-9.

Therefore, the current draft of ISO/IEC 9995-9 contains (besides many other "groups"):

- a "superscript" group, which maps the superscript version of any capital or small basic Latin letter to that letter;
- and a "subscript" group, which maps the subscript version of any small basic Latin letter to that letter, while it maps the small capital version of any capital basic Latin letter to that letter.

However, until now, these "groups" cannot be represented completely in Unicode.

This results in the following problems:

– The user cannot be served with a consistent behavior, yielding the expected result consistently for any basic Latin letter.
  While other groups like "stroked diagonally" can map the "missing" letters to combination of the letter itself + an overstriking diacritic, such cannot be done for superscripting etc.
  Access to any higher level protocol is not given, as ISO/IEC 9995-9 is not dependent of the presence of any such protocol; it maps keystrokes simply to (plain text) characters.

– Future editions of ISO/IEC 9995-9 would have to follow any single later additions to Unicode regarding superscript/subscript/smallcap Latin letters, which can only be done with a delay, and requiring changes by all devices incorporating that standard in firmware or software.

Therefore, ISO/IEC JTE1/SC35 requests to add the missing characters into Unicode and ISO/IEC 10646 at this time, thus ISO/IEC 9995-9 can rely on the standardized characters when it itself is finished through the WD/DIS/FDIS procedure, in parallel to the ISO/IEC 10646 version.

It is emphasized that such is requested only for the 2×26 basic Latin letters A...Z + a...z, and in no case for other Latin letters, or letters of other scripts. This is ensured by the very principle of ISO/IEC 9995-9, which refers explicitly to a set of 26 keys denoted by the 26 basic Latin letters.

This is in analogy to the recent decision to complete the set of squared, negative circled, and negative squared letters in the Enclosed Alphanumeric Supplement block (1F100...1F1FF).

1.3 Other advantages of having the full modifier letter series encoded

On a recent discussion on the Unicode list regarding the encoding of a subscript character which is not a letter, there were considerations that such an encoding will have its costs. E.g. in environments where such a character now is ensured to be handled by higher level protocols, it must be considered that after such an encoding, there can be the use of plain text characters in addition. This may cause development effort for the software handling this environment.

On the other hand, it was affirmed that modifier letters will be continued to be encoded whenever an appropriate (especially linguistic) use is found. It is a known fact that such new use for yet unencoded Latin modifier letters is found regularly. Thus, the mentioned development effort will raise with almost every new Unicode version.

Therefore, encoding the missing Latin modifier letters in a single package, as it is proposed here, will result in a significant advantage, as the mentioned effort will only occur one more time.
Also, it will save people a lot of time, as new proposals for single Latin modifier letters neither need to be created nor to be discussed when new plain text use is found.
The price of having encoded a limited set letters whose appropriate use is not known now (but likely to be found later anyway) is low.
Also, it may be noted that by the recent encoding of Emoji, interoperability with other standards (which is the base of this proposal) was considered as appropriate use anyway.

2. Encoding Considerations

Characters are requested to complete the following four series:

a. Small capital letters (2 of 26 are missing in Unicode 6.0: Q, X).
   (Regarding the X, it is in fact similar to the (simple) small x, but this is also the case for the already encoded U+1D0F LATIN LETTER SMALL CAPITAL O.)

b. Small superscript letters (1 of 26 is missing in Unicode 6.0: q).

c. Capital superscript letters (7 of 26 are missing: C, F, Q, S, X, Y, Z).
The superscript letters (including the single small one) are located in the last column of the Latin Extended D block, as this column already contains superscript letters.

The characters listed under a., b., and c. are placed at the end of the Latin Extended D block, to put them near other modifier letters, as well as to not conflict with other Latin letters currently proposed.

d. Small subscript letters (9 of 26 are missing: b, c, d, f, g, q, w, y, z).
These are located in the last column of the Phonetic Extensions Supplement B block, to avoid any conflicts with other proposals using this block.
Also, the possibility to split this block is allowed for, to separate a single-column block "Superscripts and Subscripts Supplement" at ABB0...ABBF, if such is considered appropriate.

3. Proposed Characters

Block: Latin Extended-D

Small capital letters

Q U+A7EE LATIN LETTER SMALL CAPITAL CAPITAL Q

X U+A7EF LATIN LETTER SMALL CAPITAL CAPITAL X

Superscript letters

C U+A7F0 MODIFIER LETTER CAPITAL C
≈ <super> 0043 C

F U+A7F1 MODIFIER LETTER CAPITAL F
≈ <super> 0046 F

Q U+A7F2 MODIFIER LETTER CAPITAL Q
≈ <super> 0051 Q

q U+A7F3 MODIFIER LETTER SMALL Q
≈ <super> 0071 q

S U+A7F4 MODIFIER LETTER CAPITAL S
≈ <super> 0053 S

X U+A7F5 MODIFIER LETTER CAPITAL X
≈ <super> 0058 X

Y U+A7F6 MODIFIER LETTER CAPITAL Y
≈ <super> 0059 Y

Z U+A7F7 MODIFIER LETTER CAPITAL Z
≈ <super> 005A Z

Block: Phonetic Extensions Supplement B

Latin subscript letters

ᵇ	U+ABB0	LATIN SUBSCRIPT SMALL LETTER B ≈ <sub> 0062 b
ᶜ	U+ABB1	LATIN SUBSCRIPT SMALL LETTER C ≈ <sub> 0063 c
ᵈ	U+ABB2	LATIN SUBSCRIPT SMALL LETTER D ≈ <sub> 0064 d
ᶠ	U+ABB3	LATIN SUBSCRIPT SMALL LETTER F ≈ <sub> 0066 f
ᵍ	U+ABB4	LATIN SUBSCRIPT SMALL LETTER G ≈ <sub> 0067 g
ᵠ	U+ABB5	LATIN SUBSCRIPT SMALL LETTER Q ≈ <sub> 0071 q
ʷ	U+ABB6	LATIN SUBSCRIPT SMALL LETTER W ≈ <sub> 0077 w
ʸ	U+ABB7	LATIN SUBSCRIPT SMALL LETTER Y ≈ <sub> 0079 y
ᶻ	U+ABB8	LATIN SUBSCRIPT SMALL LETTER Z ≈ <sub> 007A z

Properties:

```
A7EE;LATIN LETTER SMALL CAPITAL Q;Ll;0;L;;;;;N;;;;;
A7EF;LATIN LETTER SMALL CAPITAL X;Ll;0;L;;;;;N;;;;;
A7F0;MODIFIER LETTER CAPITAL C;Lm;0;L;<super> 0043;;;;N;;;;;
A7F1;MODIFIER LETTER CAPITAL F;Lm;0;L;<super> 0046;;;;N;;;;;
A7F2;MODIFIER LETTER CAPITAL Q;Lm;0;L;<super> 0051;;;;N;;;;;
A7F3;MODIFIER LETTER SMALL Q;Lm;0;L;<super> 0071;;;;N;;;;;
A7F4;MODIFIER LETTER CAPITAL S;Lm;0;L;<super> 0053;;;;N;;;;;
A7F5;MODIFIER LETTER CAPITAL X;Lm;0;L;<super> 0058;;;;N;;;;;
A7F6;MODIFIER LETTER CAPITAL Y;Lm;0;L;<super> 0059;;;;N;;;;;
A7F7;MODIFIER LETTER CAPITAL Z;Lm;0;L;<super> 005A;;;;N;;;;;
ABB0;LATIN SUBSCRIPT SMALL LETTER B;Lm;0;L;<sub> 0062;;;;N;;;;;
ABB1;LATIN SUBSCRIPT SMALL LETTER C;Lm;0;L;<sub> 0063;;;;N;;;;;
ABB2;LATIN SUBSCRIPT SMALL LETTER D;Lm;0;L;<sub> 0064;;;;N;;;;;
ABB3;LATIN SUBSCRIPT SMALL LETTER F;Lm;0;L;<sub> 0066;;;;N;;;;;
ABB4;LATIN SUBSCRIPT SMALL LETTER G;Lm;0;L;<sub> 0062;;;;N;;;;;
ABB5;LATIN SUBSCRIPT SMALL LETTER Q;Lm;0;L;<sub> 0071;;;;N;;;;;
ABB6;LATIN SUBSCRIPT SMALL LETTER W;Lm;0;L;<sub> 0077;;;;N;;;;;
ABB7;LATIN SUBSCRIPT SMALL LETTER Y;Lm;0;L;<sub> 0079;;;;N;;;;;
ABB8;LATIN SUBSCRIPT SMALL LETTER Z;Lm;0;L;<sub> 007A;;;;N;;;;;
```

4. Appendix: Complete list of small capital and modifier letters based on the basic Latin alphabet A...Z

4.1 Small capital letters A...Z:

```
A:  U+1D00 LATIN LETTER SMALL CAPITAL A
B:  U+0299 LATIN LETTER SMALL CAPITAL B
C:  U+1D04 LATIN LETTER SMALL CAPITAL C
D:  U+1D05 LATIN LETTER SMALL CAPITAL D
E:  U+1D07 LATIN LETTER SMALL CAPITAL E
F:  U+A730 LATIN LETTER SMALL CAPITAL F
G:  U+0262 LATIN LETTER SMALL CAPITAL G
H:  U+029C LATIN LETTER SMALL CAPITAL H
I:  U+026A LATIN LETTER SMALL CAPITAL I
J:  U+1D0A LATIN LETTER SMALL CAPITAL J
K:  U+1D0B LATIN LETTER SMALL CAPITAL K
L:  U+029F LATIN LETTER SMALL CAPITAL L
M:  U+1D0D LATIN LETTER SMALL CAPITAL M
N:  U+0274 LATIN LETTER SMALL CAPITAL N
O:  U+1D0F LATIN LETTER SMALL CAPITAL O
P:  U+1D18 LATIN LETTER SMALL CAPITAL P
Q:  U+A7EE LATIN LETTER SMALL CAPITAL Q  # This character is proposed here.
R:  U+0280 LATIN LETTER SMALL CAPITAL R
S:  U+A731 LATIN LETTER SMALL CAPITAL S
T:  U+1D1B LATIN LETTER SMALL CAPITAL T
U:  U+1D1C LATIN LETTER SMALL CAPITAL U
V:  U+1D20 LATIN LETTER SMALL CAPITAL V
W:  U+1D21 LATIN LETTER SMALL CAPITAL W
X:  U+A7EF LATIN LETTER SMALL CAPITAL X  # This character is proposed here.
Y:  U+028F LATIN LETTER SMALL CAPITAL Y
Z:  U+1D22 LATIN LETTER SMALL CAPITAL Z
```

4.2 Superscript capital letters A...Z:

```
A:  U+1D2C MODIFIER LETTER CAPITAL A
B:  U+1D2E MODIFIER LETTER CAPITAL B
C:  U+A7F0 MODIFIER LETTER CAPITAL C  # This character is proposed here.
D:  U+1D30 MODIFIER LETTER CAPITAL D
E:  U+1D31 MODIFIER LETTER CAPITAL E
F:  U+A7F1 MODIFIER LETTER CAPITAL F  # This character is proposed here.
G:  U+1D33 MODIFIER LETTER CAPITAL G
H:  U+1D34 MODIFIER LETTER CAPITAL H
I:  U+1D35 MODIFIER LETTER CAPITAL I
J:  U+1D36 MODIFIER LETTER CAPITAL J
K:  U+1D37 MODIFIER LETTER CAPITAL K
L:  U+1D38 MODIFIER LETTER CAPITAL L
M:  U+1D39 MODIFIER LETTER CAPITAL M
N:  U+1D3A MODIFIER LETTER CAPITAL N
O:  U+1D3C MODIFIER LETTER CAPITAL O
P:  U+1D3E MODIFIER LETTER CAPITAL P
Q:  U+A7F2 MODIFIER LETTER CAPITAL Q  # This character is proposed here.
R:  U+1D3F MODIFIER LETTER CAPITAL R
S:  U+A7F4 MODIFIER LETTER CAPITAL S  # This character is proposed here.
T:  U+1D40 MODIFIER LETTER CAPITAL T
U:  U+1D41 MODIFIER LETTER CAPITAL U
V:  U+2C7D MODIFIER LETTER CAPITAL V
W:  U+1D42 MODIFIER LETTER CAPITAL W
X:  U+A7F5 MODIFIER LETTER CAPITAL X  # This character is proposed here.
Y:  U+A7F6 MODIFIER LETTER CAPITAL Y  # This character is proposed here.
Z:  U+A7F7 MODIFIER LETTER CAPITAL Z  # This character is proposed here.
```

4.3 Superscript small letters a...z:

```
a:  U+1D43 MODIFIER LETTER SMALL A
b:  U+1D47 MODIFIER LETTER SMALL B
c:  U+1D9C MODIFIER LETTER SMALL C
d:  U+1D48 MODIFIER LETTER SMALL D
e:  U+1D49 MODIFIER LETTER SMALL E
f:  U+1DA0 MODIFIER LETTER SMALL F
g:  U+1D4D MODIFIER LETTER SMALL G
h:  U+02B0 MODIFIER LETTER SMALL H
i:  U+2071 SUPERSCRIPT LATIN SMALL LETTER I
j:  U+02B2 MODIFIER LETTER SMALL J
k:  U+1D4F MODIFIER LETTER SMALL K
l:  U+02E1 MODIFIER LETTER SMALL L
m:  U+1D50 MODIFIER LETTER SMALL M
n:  U+207F SUPERSCRIPT LATIN SMALL LETTER N
o:  U+1D52 MODIFIER LETTER SMALL O
p:  U+1D56 MODIFIER LETTER SMALL P
q:  U+A7F3 MODIFIER LETTER SMALL Q  # This character is proposed here.
r:  U+02B3 MODIFIER LETTER SMALL R
s:  U+02E2 MODIFIER LETTER SMALL S
t:  U+1D57 MODIFIER LETTER SMALL T
u:  U+1D58 MODIFIER LETTER SMALL U
v:  U+1D5B MODIFIER LETTER SMALL V
w:  U+02B7 MODIFIER LETTER SMALL W
x:  U+02E3 MODIFIER LETTER SMALL X
y:  U+02B8 MODIFIER LETTER SMALL Y
z:  U+1DBB MODIFIER LETTER SMALL Z
```

4.4 Subscript small letters a...z:

```
a:  U+2090 LATIN SUBSCRIPT SMALL LETTER A
b:  U+ABB0 LATIN SUBSCRIPT SMALL LETTER B  # This character is proposed here.
c:  U+ABB1 LATIN SUBSCRIPT SMALL LETTER C  # This character is proposed here.
d:  U+ABB2 LATIN SUBSCRIPT SMALL LETTER D  # This character is proposed here.
e:  U+2091 LATIN SUBSCRIPT SMALL LETTER E
f:  U+ABB3 LATIN SUBSCRIPT SMALL LETTER F  # This character is proposed here.
g:  U+ABB4 LATIN SUBSCRIPT SMALL LETTER G  # This character is proposed here.
h:  U+2095 LATIN SUBSCRIPT SMALL LETTER H
i:  U+1D62 LATIN SUBSCRIPT SMALL LETTER I
j:  U+2C7C LATIN SUBSCRIPT SMALL LETTER J
k:  U+2096 LATIN SUBSCRIPT SMALL LETTER K
l:  U+2097 LATIN SUBSCRIPT SMALL LETTER L
m:  U+2098 LATIN SUBSCRIPT SMALL LETTER M
n:  U+2099 LATIN SUBSCRIPT SMALL LETTER N
o:  U+2092 LATIN SUBSCRIPT SMALL LETTER O
p:  U+209A LATIN SUBSCRIPT SMALL LETTER P
q:  U+ABB5 LATIN SUBSCRIPT SMALL LETTER Q  # This character is proposed here.
r:  U+1D63 LATIN SUBSCRIPT SMALL LETTER R
s:  U+209B LATIN SUBSCRIPT SMALL LETTER S
t:  U+209C LATIN SUBSCRIPT SMALL LETTER T
u:  U+1D64 LATIN SUBSCRIPT SMALL LETTER U
v:  U+1D65 LATIN SUBSCRIPT SMALL LETTER V
w:  U+ABB6 LATIN SUBSCRIPT SMALL LETTER W  # This character is proposed here.
x:  U+2093 LATIN SUBSCRIPT SMALL LETTER X
y:  U+ABB7 LATIN SUBSCRIPT SMALL LETTER Y  # This character is proposed here.
z:  U+ABB8 LATIN SUBSCRIPT SMALL LETTER Z  # This character is proposed here.
```

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html **.**
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** ***Proposal to encode missing Latin small capital and modifier letters in the UCS***
2. Requester's name: *ISO/IEC JTC1/SC35*
3. Requester type (Member body/Liaison/Individual contribution): *Liaison contribution*
4. Submission date: *2011-03-12*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *Yes*
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): *No*
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Yes*
      Name of the existing block: *Latin Extended-D; Phonetic Extensions Supplement B*
2. Number of characters in proposal: *19*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary ( ) B.1-Specialized (small collection) (X) B.2-Specialized (large collection) ( )
   C-Major extinct ( ) D-Attested extinct ( ) E-Minor extinct ( )
   F-Archaic Hieroglyphic or Ideographic ( ) G-Obscure or questionable usage symbols ( )
4. Is a repertoire including character names provided? *Yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *Yes*
   b. Are the character shapes attached in a legible form suitable for review? *Yes*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   Not necessary; glyphs are straightforward modifications (superscripting etc.) of existing characters
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *Yes*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *No*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *No*

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N3702-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *No*
   If YES, with whom?
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *Yes*
   Reference: *see text*
4. The context of use for the proposed characters (type of use; common or rare) *rare*
   Reference: *see text*
5. Are the proposed characters in current use by the user community? *Yes*
   If YES, where? Reference: *see text*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *Yes*
   If YES, is a rationale provided? *Yes*
   If YES, reference: *to keep them in line with similar characters*
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *partially*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? *No*
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? *No*
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *No*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility character(s)? *No*
    If YES, is the equivalent corresponding unified ideographic character(s) identified?
    If YES, reference: