

Universal Multiple-Octet Coded Character Set

Doc Type: Working Group Document
Title: **Additional Characters for the UCS**
Source: Ad Hoc on Bucket 35
Status: Working Document

References: Cited separately in sections below.

Meeting Dates: 1998-09-22, 1998-09-23
Attendees: Joan Aliprand (USA)
Michael Everson (Ireland)
Klaas Ruppel (Finland)
Johan van Wingen (Netherlands)
Ken Whistler (USA)
Chris White (British Library)

Synopsis

This document constitutes the meeting report of the ad hoc committee on “Bucket 35”, the collection of characters from proposals for small additions of various characters.

From a procedural point of view, this collection is divided into two parts. The first part consists of a large number of characters from many TC46 standards proposed for encoding in 10646, along with 10 Cyrillic Sami characters discussed in the same set of documents. The second part consists of many small, unrelated proposals.

Characters Derived from TC46 Standards; Cyrillic Sami Characters

This collection of characters is dealt with in the following documents. The original TC46-related proposals are WG2 N 1741, N 1743, N 1744, N 1745, N 1746, N 1747, N 1748, N 1749. The Cyrillic Sami character proposal is N 1813 (there are earlier versions, but those documents are superseded by N 1813). WG2 N 1840 is the consolidated response of the U.S. national body to WG2 N 1741, 1743 - 1749. WG2 N 1885 is the consolidated response of the U.S. national body to WG2 N 1742, N 1813, and to those characters in several others of the TC46-related proposals which do not in fact derive from the TC46 standards. WG2 N 1847 is the consolidated response of the Irish national body to L2/98-292 (= WG2 N 1840) and to N 1885. The actual citations of particular characters are scattered all through this particular interrelated set of documents.

It is the opinion of the ad hoc that the best approach for WG2 to deal with these documents is to respond to the last in the sequence (N 1847). Any decision made on that basis of that document renders all the earlier documents moot. WG2 can then invite the Irish national body to develop superseding new proposals (if it so desires) regarding any remaining characters not satisfactorily dealt with by the response to N 1847.

Because a large number of the proposed characters proved to be controversial, they are divided here into three categories, as determined by the ad hoc: 1. Those that should be encoded, 2. Those that should not be encoded (for various reasons), and 3. Those that require further study. For those that the ad hoc determined

should be encoded, a proposed encoding point and name are provided, for use in development of an amendment to add them to 10646.

The following 28 characters from the collections mentioned above should be added to the UCS:

0222 LATIN CAPITAL LETTER OU
0223 LATIN SMALL LETTER OU
0224 LATIN CAPITAL LETTER Z WITH HOOK
0225 LATIN SMALL LETTER Z WITH HOOK
03D7 GREEK KAI SYMBOL
03D8 GREEK SMALL LETTER STIGMA
03DD GREEK SMALL LETTER DIGAMMA
03DF GREEK SMALL LETTER KOPPA
03E1 GREEK SMALL LETTER SAMPI
0488 CYRILLIC HUNDRED THOUSANDS SIGN
0489 CYRILLIC MILLIONS SIGN
048E CYRILLIC CAPITAL LETTER ER WITH TICK
048F CYRILLIC SMALL LETTER ER WITH TICK
04C5 CYRILLIC CAPITAL LETTER EL WITH DESCENDER
04C6 CYRILLIC SMALL LETTER EL WITH DESCENDER
04C9 CYRILLIC CAPITAL LETTER SHORT I WITH DESCENDER
04CA CYRILLIC SMALL LETTER SHORT I WITH DESCENDER
04CD CYRILLIC CAPITAL LETTER EM WITH DESCENDER
04CE CYRILLIC SMALL LETTER EM WITH DESCENDER
04EC CYRILLIC CAPITAL LETTER E WITH DIAERESIS
04ED CYRILLIC SMALL LETTER E WITH DIAERESIS
204A TIRONIAN SIGN ET
204B REVERSED PILCROW SIGN
204C BLACK LEFTWARDS BULLET
204D BLACK RIGHTWARDS BULLET
2139 LATIN CAPITAL LETTER ROTATED Q
2183 ROMAN NUMERAL REVERSED ONE HUNDRED
2619 REVERSED ROTATED FLORAL HEART BULLET

The following characters from the collections mentioned above should **not** be added to the UCS. The reasons are various: erroneous mapping proposed, character representable otherwise, etc.

GREEK KAI SYMBOL WITH VARIA
LATIN CAPITAL LETTER YR
LATIN SMALL LETTER YR
VECTOR OR SUM
VECTOR PRODUCT
SUM OR UNION OF CLASSES OR SETS
PRODUCT OF INTERSECTION OF CLASSES OR SETS [sic]
COMBINING RIGHT DESCENDER
COMBINING LEFT DESCENDER
SIX-SPOKED ASTERISK

The following characters from the collections mentioned above require more study:

IS INCLUDED IN SET
INCLUDES IN SET
CYRILLIC TEN THOUSANDS SIGN
SEXTILE
GREEK CAPITAL LIGATURE OU

GREEK SMALL LIGATURE OU
COMBINING LATIN SMALL LETTER A ABOVE
COMBINING LATIN SMALL LETTER E ABOVE
COMBINING LATIN SMALL LETTER R ABOVE
COMBINING LATIN SMALL LETTER Z ABOVE
COMBINING DOUBLE CARON
COMBINING DOUBLE CIRCUMFLEX
LATIN CONTRACTION REVERSED US
LATIN CONTRACTION IS
LATIN CONTRACTION SMALL IS
LATIN CONTRACTION UM
REVERSED SECTION SIGN
(all proposed Hebrew cantillation marks from N 1749)
(all other proposed Cyrillic letters from N 1744)
(all other Latin contractions from N 1747)

The Ad Hoc Committee additionally suggests the addition of the following 38 characters from various sources, as noted.

From N 1322 (Livonian characters):

022A LATIN CAPITAL LETTER O WITH DIAERESIS AND MACRON
022B LATIN SMALL LETTER O WITH DIAERESIS AND MACRON
022C LATIN CAPITAL LETTER O WITH TILDE AND MACRON
022D LATIN SMALL LETTER O WITH TILDE AND MACRON
022E LATIN CAPITAL LETTER O WITH DOT ABOVE
022F LATIN SMALL LETTER O WITH DOT ABOVE
0230 LATIN CAPITAL LETTER O WITH DOT ABOVE AND MACRON
0231 LATIN SMALL LETTER O WITH DOT ABOVE AND MACRON
0232 LATIN CAPITAL LETTER Y WITH MACRON
0233 LATIN SMALL LETTER Y WITH MACRON

From N 1817 (Nenets character):

02EE MODIFIER LETTER DOUBLE APOSTROPHE

From N 1812 (Swedish accent):

02DF MODIFIER LETTER CROSS ACCENT

From N 1838 (binary completion characters):

0226 LATIN CAPITAL LETTER A WITH DOT ABOVE
0227 LATIN SMALL LETTER A WITH DOT ABOVE
0228 LATIN CAPITAL LETTER E WITH CEDILLA
0229 LATIN SMALL LETTER E WITH CEDILLA

From N 1857 (Mongolian currency sign):

20AE TUGRIK SIGN

From N 1845 (IPA characters for disturbed speech):

02A9 LATIN SMALL LETTER FENG
02AA LATIN SMALL LETTER LES
02AB LATIN SMALL LETTER LEZ
02AC LATIN LETTER BILABIAL PERCUSSIVE
02AD LATIN LETTER BIDENTAL PERCUSSIVE
02EC MODIFIER LETTER VOICING
02ED MODIFIER LETTER UNASPIRATED
0346 COMBINING BRIDGE ABOVE
0347 COMBINING EQUALS SIGN BELOW
0348 COMBINING DOUBLE VERTICAL LINE BELOW
0349 COMBINING LEFT ANGLE BELOW
034A COMBINING NOT TILDE ABOVE
034B COMBINING HOMOTHETIC ABOVE
034C COMBINING ALMOST EQUAL TO ABOVE
034D COMBINING LEFT RIGHT ARROW BELOW
034E COMBINING UPWARDS ARROW BELOW
0362 COMBINING DOUBLE RIGHTWARDS ARROW BELOW

From N 1882 (Interlinear annotation characters)

FFF9 INTERLINEAR ANNOTATION ANCHOR
FFFA INTERLINEAR ANNOTATION SEPARATOR
FFFB INTERLINEAR ANNOTATION TERMINATOR

From N 1886

204F SOFT SPACE