ISO/IEC JTC1/SC2/WG2 N 2921

Date: **2005-01-25**

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

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

Title: **Named UCS Sequence Identifiers**

Source: **US national Body**
Project: JTC1 02.18
Status: For review by WG2
Date: 2005-01-25
Distribution: WG2
Reference:
Medium:

Overview

The US national body is requesting to amend ISO/IEC 10646 to include Rules for inclusion of Named UCS Sequence Identifiers (USI). The purpose is to specify sequences of characters that may be treated as single units, either in particular types of processing, in reference by standards, or in listing of repertoires (such as for fonts or keyboards).

Some standards, notably those developed by ISO/IEC JTC1/SC2 have a long standing tradition of using the formal name of a character as the means to identify corresponding characters across standards. With ISO/IEC 10646 as the universal character set, this practice has largely given way to using the code point in ISO/IEC 10646 as the unique identifier. However, some standards contain entities or characters that are mapped to a sequence of characters rather than to a single UCS code point. In these instances it is convenient to have a name for the sequence.

Here are hypothetical examples of such characters, and their representation as a sequence of code points.

Table 1 Examples of Named Sequences

Appearance	Code Points	Name or Linguistic Usage
ch	0063 0068	Slovak, traditional Spanish
tʰ	0074 02B0	Native American languages
x̣	0078 0323	
ƛ́	019B 0313	
ą	00E1 0328	Lithuanian
í	0069 0307 0301	
ㇰ゜	30C8 309A	Ainu in kana transcription
◌ុ	17BB 17C6	khmer vowel sign srak om

◌ំ	17B6 17C6	khmer vowel sign srak am
◌្ក	17D2 1780	khmer consonant sign coeng ka

While additional characters may be added to ISO/IEC 10646 to represent these sequences, it is not always possible to do so. One of the main reason is the normalization form C (NFC) which requires to represent a character sequence as the shortest form possible using ISO/IEC 10646-1:2000 and ISO/IEC-2:2001 repertoire. Even if a shorter form is added later, the NFC normalization process would eliminate it.

Named USIs allow the definition of sequences that can be used for mapping purposes, writing system references, and many other purposes where a character name would be used.

Notation

ISO/IEC 10646 already has a formal definition for USI in sub-clause 6.6.

Names

Names of Named USI sequences are unique within their own space and the character name space. Where possible, they are constructed by appending the names of the constituent elements together while eliding duplicate elements. Should this process result in a name that already exists, the name is modified suitably to guarantee uniqueness among Named USI and individual character names.

Table 2 Examples of hypothetical sequence names

USI	Alternate representation of sequence	Name
<0041, 0043, 0043>	<A, B, C>	LATIN CAPITAL LETTER A B C
<00CA, 0046>	<AE, F>	LATIN CAPITAL LETTER AE F
<005X, xxxx>	<X, COMBINING DIACRITIC ABOVE>	LATIN CAPITAL LETTER X WITH DIACRITIC ABOVE

In all cases the name should follow the rules provided in Annex L in ISO/IEC 10646.

Content

The file at http://www.unicode.org/Public/4.1.0/ucd/NamedSequences-4.1.0d5.txt contains the current set.

Implication for ISO/IEC 10646

A new clause needs to be added, preferably between clause 28 and 29 to cover Named USIs specifying the concepts expressed above. That clause may contain the list of the currently proposed Named USIs or simply link to it. The propose text is as follows:

29 Named UCS Sequence Identifiers

A named UCS Sequence Identifier (USI) is a USI associated to a name following the same construction rules as character names. These rules are given in Annex L. The uniqueness rule expressed in Annex L applies to both character names and named USIs considered as a whole.

NOTE 1 – Where possible, the names of these USIs are constructed by appending the names of the constituent elements together while eliding duplicate elements. Should this process result in a name that already exists, the name is modified suitably to guarantee uniqueness.

NOTE 2 – The purpose of these named USIs is to specify sequences of characters that may be treated as single units, either in particular types of processing, in reference by standards, in listing of repertoires (such as for fonts or keyboards).

The following list provides a description of these named UCS Sequence Identifiers.

USI USI name

<0100, 0300>	LATIN CAPITAL LETTER A WITH MACRON AND GRAVE
<0101, 0300>	LATIN SMALL LETTER A WITH MACRON AND GRAVE
<00E1, 0328>	LATIN SMALL LETTER A WITH ACUTE AND OGONEK
<0045, 0329>	LATIN CAPITAL LETTER E WITH VERTICAL LINE BELOW
<0065, 0329>	LATIN SMALL LETTER E WITH VERTICAL LINE BELOW
<00C8, 0329>	LATIN CAPITAL LETTER E WITH VERTICAL LINE BELOW AND GRAVE
<00E8, 0329>	LATIN SMALL LETTER E WITH VERTICAL LINE BELOW AND GRAVE
<00C9, 0329>	LATIN CAPITAL LETTER E WITH VERTICAL LINE BELOW AND ACUTE
<00E9, 0329>	LATIN SMALL LETTER E WITH VERTICAL LINE BELOW AND ACUTE
<00CA, 0304>	LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND MACRON
<00EA, 0304>	LATIN SMALL LETTER E WITH CIRCUMFLEX AND MACRON
<00CA, 030C>	LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND CARON
<00EA, 030C>	LATIN SMALL LETTER E WITH CIRCUMFLEX AND CARON
<012A, 0300>	LATIN CAPITAL LETTER I WITH MACRON AND GRAVE
<012B, 0300>	LATIN SMALL LETTER I WITH MACRON AND GRAVE
<0069, 0307, 0301>	LATIN SMALL LETTER I WITH DOT ABOVE AND ACUTE
<006E, 0360, 0067>	LATIN SMALL LETTER NG WITH TILDE ABOVE
<004F, 0329>	LATIN CAPITAL LETTER O WITH VERTICAL LINE BELOW
<006F, 0329>	LATIN SMALL LETTER O WITH VERTICAL LINE BELOW
<00D2, 0329>	LATIN CAPITAL LETTER O WITH VERTICAL LINE BELOW AND GRAVE
<00F2, 0329>	LATIN SMALL LETTER O WITH VERTICAL LINE BELOW AND GRAVE
<00D3, 0329>	LATIN CAPITAL LETTER O WITH VERTICAL LINE BELOW AND ACUTE
<00F3, 0329>	LATIN SMALL LETTER O WITH VERTICAL LINE BELOW AND ACUTE
<0053, 0329>	LATIN CAPITAL LETTER S WITH VERTICAL LINE BELOW
<0073, 0329>	LATIN SMALL LETTER S WITH VERTICAL LINE BELOW
<016A, 0300>	LATIN CAPITAL LETTER U WITH MACRON AND GRAVE
<016B, 0300>	LATIN SMALL LETTER U WITH MACRON AND GRAVE
<10E3, 0302>	GEORGIAN LETTER U-BRJGU
<17D2, 1780>	KHMER CONSONANT SIGN COENG KA
<17D2, 1781>	KHMER CONSONANT SIGN COENG KHA
<17D2, 1782>	KHMER CONSONANT SIGN COENG KO
<17D2, 1783>	KHMER CONSONANT SIGN COENG KHO
<17D2, 1784>	KHMER CONSONANT SIGN COENG NGO
<17D2, 1785>	KHMER CONSONANT SIGN COENG CA
<17D2, 1786>	KHMER CONSONANT SIGN COENG CHA
<17D2, 1787>	KHMER CONSONANT SIGN COENG CO
<17D2, 1788>	KHMER CONSONANT SIGN COENG CHO
<17D2, 1789>	KHMER CONSONANT SIGN COENG NYO
<17D2, 178A>	KHMER CONSONANT SIGN COENG DA
<17D2, 178B>	KHMER CONSONANT SIGN COENG TTHA
<17D2, 178C>	KHMER CONSONANT SIGN COENG DO
<17D2, 178D>	KHMER CONSONANT SIGN COENG TTHO
<17D2, 178E>	KHMER CONSONANT SIGN COENG NA
<17D2, 178F>	KHMER CONSONANT SIGN COENG TA
<17D2, 1790>	KHMER CONSONANT SIGN COENG THA
<17D2, 1791>	KHMER CONSONANT SIGN COENG TO
<17D2, 1792>	KHMER CONSONANT SIGN COENG THO
<17D2, 1793>	KHMER CONSONANT SIGN COENG NO
<17D2, 1794>	KHMER CONSONANT SIGN COENG BA
<17D2, 1795>	KHMER CONSONANT SIGN COENG PHA
<17D2, 1796>	KHMER CONSONANT SIGN COENG PO
<17D2, 1797>	KHMER CONSONANT SIGN COENG PHO
<17D2, 1798>	KHMER CONSONANT SIGN COENG MO
<17D2, 1799>	KHMER CONSONANT SIGN COENG YO
<17D2, 179A>	KHMER CONSONANT SIGN COENG RO
<17D2, 179B>	KHMER CONSONANT SIGN COENG LO
<17D2, 179C>	KHMER CONSONANT SIGN COENG VO
<17D2, 179D>	KHMER CONSONANT SIGN COENG SHA
<17D2, 179E>	KHMER CONSONANT SIGN COENG SSA
<17D2, 179F>	KHMER CONSONANT SIGN COENG SA
<17D2, 17A0>	KHMER CONSONANT SIGN COENG HA
<17D2, 17A1>	KHMER CONSONANT SIGN COENG LA
<17D2, 17A2>	KHMER VOWEL SIGN COENG QA
<17D2, 17A7>	KHMER INDEPENDENT VOWEL SIGN COENG QU
<17D2, 17AB>	KHMER INDEPENDENT VOWEL SIGN COENG RY
<17D2, 17AC>	KHMER INDEPENDENT VOWEL SIGN COENG RYY
<17D2, 17AF>	KHMER INDEPENDENT VOWEL SIGN COENG QE
<17BB 17C6>	KHMER VOWEL SIGN OM
<17B6, 17C6>	KHMER VOWEL SIGN AAM
<31F7, 309A>	KATAKANA LETTER AINU P
<02E5, 02E9>	MODIFIER LETTER EXTRA-HIGH EXTRA-LOW CONTOUR TONE BAR

All the allowed Named USIs are defined in this clause; all other such sequences are undefined.

Annex L needs to be updated to cover Named USIs. This will be addressed at the same time that the Annex is updated to take in account block names. The main editing is to replace the terms ‘character name’ and ‘name of character’ by a more generic term.
