ISO/IEC JTC1/SC2/WG2 N3184
L2/06-357
2006-10-30

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

Doc Type: **Working Group Document**
Title: **On CYRILLIC LETTER OMEGA WITH TITLO and on CYRILLIC LETTER UK**
Source: **Michael Everson, David Birnbaum (University of Pittsburgh), Ralph Cleminson (University of Portsmouth), Ivan Derzhanski (Bulgarian Academy of Sciences), Vladislav Dorosh (irmologion.ru), Alexej Kryukov (Moscow State University), and Sorin Paliga (University of Bucharest)**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2006-10-30**

1. Introduction. The Unicode Technical Committee has recently discussed problems which users of the Cyrillic block have had with Ѽ U+047C CYRILLIC CAPITAL LETTER OMEGA WITH TITLO and ѽ U+047D CYRILLIC SMALL LETTER OMEGA WITH TITLO. Recent discussion about adding a number of missing Cyrillic characters has turned up another difficulty regarding Oy|Ꙋ U+0478 CYRILLIC CAPITAL LETTER UK and oy|ꙋ U+0479 CYRILLIC SMALL LETTER UK, which also needs consideration and resolution. This paper discusses these problems and proposes solutions for them.

2. The problem of CYRILLIC LETTER OMEGA WITH TITLO. The chief difficulty is that this character appears to be badly misnamed, which means that it has become of no use to anyone. The real sequence ѡ U+0461 CYRILLIC SMALL LETTER OMEGA + ◌҃ U+0483 COMBINING CYRILLIC TITLO is *not* equivalent to ѽ U+047D CYRILLIC SMALL LETTER OMEGA WITH TITLO, and so there is clearly scope for multiple spellings. Moreover, there is no justification for having a precomposed form for *omega with titlo*. Such a character has no particularly special usefulness in Cyrillic to warrant its unique encoding. Professor Ralph Cleminson of the University of Portsmouth proposed to the UTC in January 2006 (L2/06-011) that the glyph for these characters be changed to reflect the only letters that seemed to make sense: ᘐ *capital letter beautiful omega* and ᘑ *small letter beautiful omega*. The "beautiful omega" (*красная омега*) is used in exclamations like "O!", and, as Deborah Anderson noted in L2/06-292, if the code position had been intended to represent this character, it "would account for there being no decomposition into omega + titlo". After a period of public review (L2/06-033), the UTC made the decision *not* to change the glyph on 2006-08-22:

> Cyrillic Omega with Titlo is left unchanged. The assessment of the UTC is that changing the glyph to be the glyph for "beautiful omega" is inappropriate, and any such character should be encoded separately.

It is not easy to see why this is "inappropriate", given the fact that in Unicode 1.0 and the first edition of ISO/IEC 10646, the glyph shown is clearly that of the "beautiful omega":

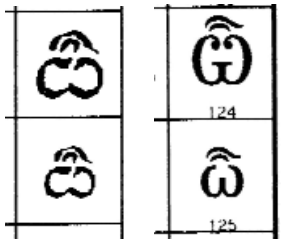

Figure 1. The OMEGA WITH TITLO from Unicode 1.0 and the first edition of ISO/IEC 10646-1.

It must be the case that the glyphs were changed for Unicode 2.0 on the basis of the character name, but that change has only made the character pretty much unusable to anyone. It can certainly not be recommended for use for *omega* with *titlo*, since that is correctly represented with ѡ U+0461 CYRILLIC SMALL LETTER OMEGA + ◌҃ U+0483 COMBINING CYRILLIC TITLO just as any letter with *titlo* is.

2a. Solution 1 for CYRILLIC LETTER OMEGA WITH TITLO. One solution would be to do more or less as was proposed previously by Professor Cleminson: change the glyphs from Ѿ and ѿ and give two notes.

047C Ҁ CYRILLIC CAPITAL LETTER OMEGA WITH TITLO
047D ҁ CYRILLIC SMALL LETTER OMEGA WITH TITLO
= cyrillic "beautiful omega"
• despite its character name, this letter does not have a titlo

One problem with this is that the "broad omega" used for this letter is needed as a stand-alone character as well as a base character for other diacritics *besides* ◌҆ U+0486 COMBINING CYRILLIC PSILI PNEUMATA and ◌̑ U+0311 COMBINING INVERTED BREVE. So if Solution 2a is accepted then two additional characters will also need to be added:

051x Ԍ CYRILLIC CAPITAL LETTER BROAD OMEGA
051x ԍ CYRILLIC SMALL LETTER BROAD OMEGA

Another problem: with these two characters, *again* we would have ambiguous spelling, because "beautiful omega" could be represented either by ҁ U+047D or by ԍ U+051x + ◌҆ U+0486 and ◌̑ U+0311, and those two representations would *not* be canonically equivalent.

In favour of Solution 2a is the fact that a number of ParaGraph/ParaType fonts (**www.paratype.com**), as well as the TEX-derived Computer Modern Unicode fonts (**cm-unicode.sourceforge.net**) have the Unicode 1.0 glyphs, and are used fairly widely throughout Cyrillia. If Solution 2a is chosen, a third note might help discourage ambiguous spellings with BROAD OMEGA, though it could not *prevent* them:

• this character is not decomposable into a base with diacritics

2b. Solution 2 for CYRILLIC LETTER OMEGA WITH TITLO. A simpler solution which allows users to write both "beautiful omega" as well as other letters—*without orthographic ambiguity*—would change the glyphs from Ѿ and ѿ and give three notes.

047C Ԍ CYRILLIC CAPITAL LETTER OMEGA WITH TITLO
047D ԍ CYRILLIC SMALL LETTER OMEGA WITH TITLO
= cyrillic "broad omega"
• used with U+0486 and U+0311 for "beautiful omega"
• despite its character name, this letter is a base character with no diacritic

The main argument in favour of Solution 2b is that it is simple and allows no possibility of multiple spellings.

In any case, given the clear evidence shown in Unicode 1.0 and the first edition of ISO/IEC 10646, it appears that the status quo (retaining the Unicode 2.0 glyphs for U+047C and U+047D) is not tenable. It is certainly not preferred by the Slavicists who are expected to use the characters. *Any solution (whether 2a or 2b) is better than keeping the character as it is*.

3. The problem of CYRILLIC LETTER UK. The root of the problem here is that the sound [u] is written in different ways at different points in the history of Cyrillic. Early on, two conventions were used: <оу> and <ѹ>, which were both called by the name *uk*. (These derive from Greek practice of writing [u] with omicron and upsilon, and ligating the two vertically.) Along with these, the letter <у> U+0443 CYRILLIC SMALL LETTER U was also used for [u] as it is in modern Russian. In typical discussions of the Slavonic alphabet, <Оу>/<оу> and <Ѹ>/<ѹ> are both listed together on the same line next to the name *uk*. It is possible to distinguish these in modern parlance by calling the first *digraph uk* and the second *monograph uk*.

The UCS character name CYRILLIC LETTER UK does not distinguish between *digraph uk* and *monograph uk*. Although the code charts show <Оу> and <оу> shapes, UCS-encoded fonts do exist which do not use the *digraph uk* glyphs, and because glyphs are informative and "LETTER UK" is not specific, it cannot be said that those fonts are "non-conformant". Examples of these fonts are:

- *Lucida Grande* (Apple's system font): 𝟾, 𝟾
- *DejaVu Sans* (Linux Kubuntu distribution): 𝟾, 𝟾
- *CMU Sans Serif* (TEX's Computer Modern Unicode font): 𝟿, 𝟾
- *CMU Serif* (TEX's Computer Modern Unicode font): 𝟿, 𝟾

UCS-encoded fonts which do use the *digraph uk* glyph are actually rather rare. The font *FreeSerif* (Linux Kubuntu distribution) uses the shapes <ОУ> and <оу>—the former being, arguably, non-conformant, depending on one's interpretation of what "CAPITAL LETTER UK" might mean. ParaType's *Newton* font also uses the shapes <ОУ> and <оу>. On Windows XP, the overwhelming majority of fonts with Cyrillic characters do not support the range U+0460-047F at all. The only one I found which does is *MS UI Gothic*, which gives glyphs <Оу> and <оу>. The *SIL Doulos* font also does not support the range U+0460-047F (in its 2004 version 4.001 or its 2006 version 4.014).

UCS-encoded fonts with Slavonic letterforms are also difficult to find. The "experimental" *Kirillica Nova Unicode* font is on a web page that hasn't been updated since 1999. It "encodes" <оу> as U+0479 (UK), <ѹ> as U+0443 (CYRILLIC U), and Cyrillic <у> as U+0079 (LATIN Y). Most Slavonic fonts easily accessible on the web are 8-bit glyph replacement fonts based on ASCII keyboard input. The **irmologion.ru** site offers eight different and very high-quality fonts. In these, <оу> is encoded as U+0075 (LATIN U), <ѹ> as U+0443 (CYRILLIC U), and Cyrillic <у> as U+00B5 (MU); U+0479 (UK) is not used.

In short, the treatment of these Cyrillic letters is underspecified in ISO/IEC 10646 and the Unicode Standard. Two additional points need to be made before we can proceed to the discussion of solutions. The Unicode Standard says nothing about ꙗ *yotated a*; one may presume that it is to be unified with я U+044F YA. This would be incorrect, because я derives historically from ѧ U+0467 LITTLE YUS; a proposal to encode the missing ꙗ *yotated a* is forthcoming, but it must be noted that the overwhelming majority of current Slavonic fonts are constrained to use U+044F (though the <я> glyph is used in those intended for non-linguists). Texts encoding ꙗ as U+044F will have to be transcoded when the new U+051x is added. The second point is related to the Unicode position that "basic Old Cyrillic uk is unified with CYRILLIC LETTER U". This also does not work, in part because manuscripts are known to use both of them contrastively, and in part because while у is used with a numeric value, ѹ is only used as a letter. The two different *uk*s, ѹ and оу, are themselves used contrastively in standard orthographies.

What we have regarding CYRILLIC LETTER UK is quite a mess. It cannot be said that there is a body of "conformant" text that can survive changes to Cyrillic in the UCS unscathed; *everyone* will have to transcode their texts one way or another in order to converge on common practice.

One additional problem has to do with the interpretation of CYRILLIC LETTER UK as a unitary digraph representing two individual letters. In casing, at present only <Оу>/<ОУ> and <оу> are allowed, but this does not meet stated user requirements for upper-casing *vs* title-casing. One solution might be to add a new character <ОУ>, but since U+0478 is already the upper-case of U+0479, this could cause case-folding problems, which might mean that the glyph for U+0478 should be changed to <ОУ>, and a new U+051x CYRILLIC CAPITAL LETTER O WITH SMALL LETTER U would be added with the glyph <Оу>—adding yet more difficulty to the equation.

As a matter of security, it is not unreasonable to assume that one of the reasons that so few vendors support U+0478 and U+0479 at all is that with <Оу>/<оу> glyphs the characters are indistinguishable from О U+041E CYRILLIC CAPITAL LETTER O, о U+041E CYRILLIC SMALL LETTER O, and у U+041E CYRILLIC SMALL LETTER U. Is it any wonder that Windows XP fonts do not support CYRILLIC LETTER UK, or that those on the Mac OS and Linux which do, do so with the *monograph uk* glyph? The fact that CYRILLIC LETTER UK is unsupported is not advantageous to either the scholarly or the ecclesiastical communities for whom this character is intended.

Ultimately, however, the use of a single character for *digraph uk* is a proposition which does not really have any justification. The digraph <оу> is not like other Cyrillic "digraphs" like <ы> or the as-yet unencoded <ъı>. When letterspacing is used for emphasis, <оу> is spaced; the others do not. Compare **доуҳати** (доухати) 'to blow' with **дъıҳати** (дъıхати) 'to breathe': when letterspaced, the correct forms are **д о у ҳ а т и** (д о у х а т и) but **д ъ ı ҳ а т и** (д ъ ı х а т и).

This is analogous to the way the *digraph uk* titlecases. The user may write the the word for 'lump' in a number of different ways depending on orthography: оукроухъ, оукрѹхъ, ѹкроухъ, ѹкрѹхъ, In title-casing this should be Оукроухъ, Оукрѹхъ, Ѹкроухъ, Ѹкрѹхъ; in all-caps it should be ОУКРОУХЪ, ОУКРѸХЪ, ѸКРОУХЪ, ѸКРѸХЪ. For comparison, these are given here in a Slavonic font: normal **оукроуҳъ**, **оукрѹхъ**, **ѹкроуҳъ**, **ѹкрѹхъ**, title case **Оукроуҳъ**, **Оукрѹхъ**, **Ѹкроуҳъ**, **Ѹкрѹхъ**, and all-caps **ОУКРОУХЪ**, **ОУКРѸХЪ**, **ѸКРОУХЪ**, **ѸКРѸХЪ**.

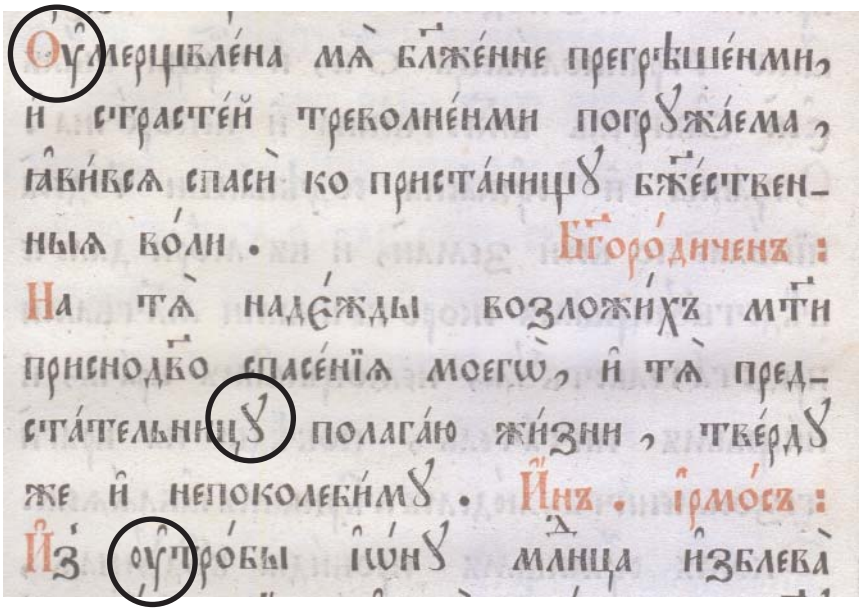

Figure 2. Example from an 1861 life of St Nicholas, showing *digraph uk* clearly distinguished as two separate characters (one red, one black), alongside *monograph uk*.

The only *uk* which makes sense for U+0479 is the *monograph uk*, and *digraph uk* should always have been considered to be a string of two characters. We offer, however, two solutions, one which retains *digraph uk* at U+0479 and one which changes it to *monograph uk*.

3a. Solution 1 for CYRILLIC LETTER UK. This solution is conservative, but disadvantageous in that it maintains ambiguity in spelling and preserves the glyphs which could be considered to be security risks. Add three characters and a note:

041E О CYRILLIC CAPITAL LETTER O
0423 У CYRILLIC CAPITAL LETTER U
043E о CYRILLIC SMALL LETTER O
0443 у CYRILLIC SMALL LETTER U
0478 Оу CYRILLIC CAPITAL LETTER UK
0479 оу CYRILLIC SMALL LETTER UK
051x Ꙋ CYRILLIC CAPITAL LETTER MONOGRAPH UK
051x ꙋ CYRILLIC SMALL LETTER MONOGRAPH UK
051x ОУ CYRILLIC CAPITAL LETTER OU
• all-caps for U+0478

Unless, with glyph change, add three characters and a note:

041E О CYRILLIC CAPITAL LETTER O
0423 У CYRILLIC CAPITAL LETTER U
043E о CYRILLIC SMALL LETTER O
0443 у CYRILLIC SMALL LETTER U
0478 ОУ CYRILLIC CAPITAL LETTER UK (*glyph change*)
0479 оу CYRILLIC SMALL LETTER UK
051x Ꙋ CYRILLIC CAPITAL LETTER MONOGRAPH UK
051x ꙋ CYRILLIC SMALL LETTER MONOGRAPH UK
051x Оу CYRILLIC CAPITAL LETTER O WITH SMALL LETTER U
• title-case for U+0478

It seems unlikely that this will lead to wider implementation of U+0478 or U+0479. Spelling ambiguity is not addressed, and security is not addressed. Another solution might be to deprecate U+0478 or U+0479 entirely, but this is also not a secure solution. Changing the glyphs, on the other hand, could *encourage* security-conscious firms to add Ꙋ and ꙋ to their fonts at these positions, in order to help *prevent* spoofing using these characters.

3b. Solution 2 for CYRILLIC LETTER UK. A simpler solution which allows users to write both *digraph uk* and *monograph uk—without the delay involved in adding new characters to the standard—*would change the glyphs from Оу and оу and give two notes.

041E О CYRILLIC CAPITAL LETTER O
0423 У CYRILLIC CAPITAL LETTER O
043E о CYRILLIC SMALL LETTER U
0443 у CYRILLIC SMALL LETTER U
0478 Ꙋ CYRILLIC CAPITAL LETTER UK (*glyph change*)
0479 ꙋ CYRILLIC SMALL LETTER UK (*glyph change*)
= “monograph uk”
• for “digraph uk” use U+043E and U+0443

It is worth noting the costs of Solution 3b.

1. It disunifies “ꙋ” from “y”, but this is inevitable (as the eventual disunification of “д” from “я” will be). Users of “y” for “ꙋ” will have to re-encode text.
2. It deprecates U+0479 for use to represent “oy” in favour of the two letters “o” and “y”. Users of U+0479 for “oy” will have to re-encode text.

It is worth noting the benefits of Solution 3b.

1. It disunifies “ꙋ” from “y”.
2. It removes U+0479 from being a security risk and allows companies to use a code point they are otherwise avoiding.
3. It leads to *unambiguous representation* of “oy”.
4. It avoids potential complaints about “duplicate encoding” of a new U+051x since “LETTER UK” correctly describes “ꙋ”.
5. It is faster, as the characters are already present in the standard.
6. It is simpler.

In any case, given the need to distinguish *digraph uk* from *monograph uk* and the clear confusion among implementers as to what U+0478 and U+0479 mean, it appears that the status quo is not tenable. It is certainly not preferred by the Slavicists who are expected to use the characters. *Any solution (whether 3a or 3b) is better than keeping the character as it is*.