L2/02-214
ISO/IEC JTC1/SC2/WG2 N2446
2002-05-10

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 for dealing with unmapped characters from ISO TC46/SC4 character sets
Source: Randall K. Barry
Status: Expert Contribution
Date: 2002-05-10

A. Administrative

1. Title
Proposal for dealing with unmapped characters from ISO TC46/SC4 character sets
2. Requester's name
Randall K. Barry
3. Requester type
Expert contribution
4. Submission date
2002-05-10
5. Requesters' reference
6a. Completion
This is a complete proposal
6b. More information to be provided?
No.

B. Technical – General

1a. New script? Name?
No.
1b. Addition of characters to existing block? Name?
Yes. Control, Latin, Hebrew, Cyrillic
2. Number of characters
54 (4 + 22 + 8 + 20)
3. Proposed category
Category A
4. Proposed level of implementation and rationale
Various levels; involves characters of differing types
5a. Character names included in proposal?
Yes.
5b. Character names in accordance with guidelines?
Yes.
5c. Character shapes reviewable?
Yes. In existing ISO standards.
6a. Who will provide computerized font?

Michael Everson, Everson Typography.

6b. Font(s) currently available?

Yes.

6c. Font format?

TrueType.

7a. Are references (to other character sets, dictionaries, descriptive texts. etc.) provided?

Yes.

7b. Are published examples (such as samples from newpapers, magazines, or other sources), of use of proposed characters attached?

No.

8. Does the proposal address other aspects of character data processing?

No.

C. Technical – Justification

1. Contact with the user community?

Yes. Randall Barry is convener of TC46/SC4/WG1, and a librarian at the Library of Congress.

2. Information on the user community?

Libraries have some of the largest databases of machine readable text in various scripts in the world.

3a. The context of use for the proposed characters?

The proposed characters are in used by implementers of the existing ISO standards from which them come.

3b. Reference

See the references to existing ISO standards in each section below.

4.a. Proposed characters in current use?

Yes, although exact number of implementers is not clear.

4b. Where?

In libraries, large and small.

5a. Characters should be encoded entirely in BMP?

Yes.

5b. Rationale

Keeping them with the other characters for each script involved.

6. Should characters be kept in a continuous range?

Code assignments within proximity of the rest of the script would be useful.

7a. Can the characters be considered a presentation form of an existing character or character sequence?

No. Any similarities are accidental.

7b. Where?

7c. Reference

8a. Can any of the characters be considered to be similar (in appearance or function) to an existing character?

No.

8b. Where?

8c. Reference

9a. Combining characters or use of composite sequences included?

Yes, in the case of four characters out of the 54 additions

9b. List of composite sequences and their corresponding glyph images provided?
No.
10. Characters with any special properties such as control function, etc. included?
Yes. Four characters.

D. Proposal

During the 30 years of its activity in the area of character sets, ISO TC46 (Information and Documentation) developed 13 International Standards within its Subcommittee 4 (Computer Applications in Libraries), Working Group 1 (Character Sets). The following International Standards were published between 1971 and 2001: ISO 5426 (Extended Latin), 5426-2 (Extended Latin for minor European languages and obsolete typography), 5427 (Extended Cyrillic), 5428 (Greek), 6438 (African languages), 6630 (Bibliographic control characters), 6861 (Glagolitic), 6862 (Mathematics), 8957 (Hebrew), 10585 (Armenian), 10586 (Georgian), 10754 (Cyrillic for non-slavic languages), and 11822 (Extended Arabic). These standards include 1,059 coded characters.

Beginning in 1991, TC46/SC4/WG1 made a concerted effort to follow and support the work of JTC1/SC2/WG2 which at that time was preparing the first edition of ISO/IEC 10646. Part of the support provided to JTC1 was documentation for the 1,059 characters in TC46 sets, with the hope that ISO/IEC 10646 would include them, thus providing a means for libraries to shift dependence from the thirteen separate character set standards to a single, universal character set. It should be noted that all thirteen TC36 character sets were developed prior to the existence of the *Character Encoding Model*. For various reasons, some of the characters in the thirteen TC46 characters sets have not incorporated into ISO/IEC 10646. Fortunately, through work by experts in both JTC1 and TC46, considerable progress has been made in accommodating the full repertoire of characters in the TC46 sets. As of May 2002, only 67 characters from TC46 sets cannot be cleanly mapped to ISO/IEC 10646. (**Note:** This tally excludes the 90 characters from ISO 6861 (Glagolitic) for which JTC1 is already considering a new repertoire of characters.)

This proposal suggests a way to deal with the 67 characters from ISO TC46 character sets which cannot be mapped cleanly to ISO/IEC 10646. The characters involved include 22 characters from ISO 5426-2 (Extended Latin for minor European languages and obsolete typography), 15 characters from ISO 6630 (Bibliographic control characters), 8 characters from ISO 8957 (Hebrew), and 22 characters from ISO 10754 (Extended Cyrillic for non-Slavic languages). It should be noted that only a small number of control characters from ISO 6630 are of critical importance to libraries at this time. The implementation status in libraries for the remaining control and graphic characters detailed below is not clear, since these TC46 character sets have enjoyed limited implementation at best since their publication. Thus, the bibliographic control characters will be dealt with first and should be considered of the highest importance to libraries. In all, the addition of only 54 new characters is suggested by this proposal.

ISO 6630 - Bibliographic Control Characters

ISO 6630 defines a small repertoire of 15 control characters for use in bibliographic data to

support specialized processing the data. Four processes are accommodated by the control characters defined. They are:filing, formatting, handling of annotations, and indexing. The set was defined as an extension of ISO 646, which included another small repertoire of control characters, six of which are used in bibliographic data. The full repertoire of 8-bit characters in ISO 6630 is as follows:

Hex	*Character Name*	*Control Type*
87	CLOSE-UP FOR SORTING (CUS)	filing
88	NON-SORTING CHARACTER(S), BEGIN (NSB)	filing
89	NON-SORTING CHARACTER(S), END (NSE)	filing
8B	PARTIAL LINE DOWN (PLD)	formatting
8C	PARTIAL LINE UP (PLU)	formatting
91	EMBEDDED ANNOTATION, BEGIN (EAB)	annotation
92	EMBEDDED ANNOTATION, END (EAE)	annotation
95	SORTING INTERPOLATION, BEGIN (SIB)	filing
96	SORTING INTERPOLATION, END (SIE)	filing
97	SECONDARY SORTING VALUE, BEGIN (SSB)	filing
98	SECONDARY SORTING VALUE, END (SSE)	filing
9C	KEY-WORD, BEGINNING (KWB)	indexing
9D	KEY-WORD, END (KWE)	indexing
9E	PERMUTATION STRING, BEGIN (PSB)	indexing
9F	PERMUTATION STRING, END (PSE)	indexing

Since its publication in 1986, libraries have widely implemented only four of the 15 bibliographic control characters defined. The four widely used characters are:

Hex	*Character Name*	*Control Type*
88	NON-SORTING CHARACTER(S), BEGIN (NSB)	filing
89	NON-SORTING CHARACTER(S), END (NSE)	filing
8B	PARTIAL LINE DOWN (PLD)	formatting
8C	PARTIAL LINE UP (PLU)	formatting

The first two control characters, NON-SORTING CHARACTER(S), BEGIN (NSB) and NON-SORTING CHARACTER(S), END (NSE) are defined for use in the MARC 21 and UNIMARC formats. They are also used in other formats related to these popular MARC implementations. UNIMARC implementations have been using these control characters for decades. Their use in MARC 21 was approved recently as a replacement for a less elegant technique for delimiting non-sorting characters in bibliographic data. The second pair of control characters: PARTIAL LINE DOWN (PLD) and PARTIAL LINE UP (PLU) are used in UNIMARC and related MARC implementations to encode super and subscript characters that occur in bibliographic data.

ISO/IEC 10646 and Unicode (its industry subset) both accommodate, minimally, a set of extended control characters by reserving character codes 0080 to 009F. Although these character codes are often associated with the “C1" control character set from ISO 6429, the standard leaves their interpretation up to the application processing the data. The pair of characters from ISO 6630, character 8B (Partial line down) and 8C (Partial line up) could be clearly mapped to

ISO/IEC 10646 characters 008B and 008C, respectively, if their definitions were more solid. The second pair of ISO 6630 characters needed by libraries, namely character 88 (Nonsorting characters begin) and hex 89 (Nonsorting characters end) could be mapped to 0098 (Start of string) and 009C (String terminator), respectively. The ISO 6429 control characters were defined more loosely in terms of their actual implementation, but would seem to meet libraries' needs for sorting applications, if their definitions were more solid in ISO/IEC 10646. Libraries would like to migrate their bibliographic data to ISO/IEC 10646 without loss of information, particularly with regard to encodings that currently isolate nonfiling strings and super or subscript characters.

The mapping of the other 11 bibliographic control characters from ISO 6630 is not as serious a matter to the libraries. These additional control characters do not appear to have been widely implemented, if at all. Although their functionality is needed by libraries, work-arounds have been developed already over the years that rely on other graphic characters and application-level solutions to handle things such as annotations and interpolations. It should be noted here, however, that the general need to handle annotations and interpolations in text could be supported by control characters such as those defined in ISO 6630. It has been suggested that rubies, kana string interpolations often associated with kanji characters in Japanese text, could benefit from the existence of control characters such as Embedded Annotation, Begin (EAB) and Embedded Annotation, End (EAE). In spite of that potential use of additional ISO 6630 characters, TC46/SC4 will only press for solid mappings for the four mentioned above. It is clear that lacking clearer need for the others, a case for their inclusion in ISO/IEC 10646 is hard to make.

ISO 5426-2 - Extended Latin for minor European languages and obsolete typography

Of the 70 graphic characters defined in the ISO 5426-2, the extended Latin character set for minor European languages and obsolete typography, only 22 have not been mapped to characters in ISO/IEC 10646. They are as follows:

Hex	*Character Name*	*Meaning*
2A	CONTRACTION LATIN SMALL LETTER REVERSED C WITH OGONEK	LATIN CONTRACTION CON
2B	CONTRACTION MARK WAVY DIGIT FOUR	LATIN CONTRACTION RUM
2C	REVERSED SECTION SIGN	LATIN CONTRACTION ORUM
2D	CONTRACTION LATIN SMALL LETTER REVERSED SCRIPT E	LATIN CONTRACTION US
39	CONTRACTION MARK WAVY DIGIT SEVEN WITH MIDDLE TILDE	LATIN CONTRACTION ET
3A	CONTRACTION MARK LATIN SMALL LETTER C WITH TAIL	LATIN CONTRACTION CO
3E	CONTRACTION MARK DOTLESS QUESTION MARK	LATIN CONTRACTION ER
42	COMBINING DOUBLE CARON	LATIN CONTRACTION
43	COMBINING DOUBLE CIRCUMFLEX	LATIN CONTRACTION
44	COMBINING GRAVE AND CIRCUMFLEX	LATIN CONTRACTION
48	COMBINING LATIN SMALL LETTER Z ABOVE	LATIN CONTRACTION
65	LATIN CAPITAL LETTER P WITH MIDDLE TILDE	LATIN CAPITAL CONTRACTION PRO
66	LATIN CAPITAL LETTER P WITH BELT	LATIN CAPITAL CONTRACTION PRO WITH BELT
67	LATIN CAPITAL LETTER P WITH STROKE	LATIN CAPTIAL CONTRACTION PER

68	LATIN CAPITAL LETTER Q WITH STROKE	LATIN CAPITAL CONTRACTION QUO
6D	LATIN SMALL LETTER QP	LATIN CAPITAL CONTRACTION QUI
75	LATIN SMALL LETTER P WITH MIDDLE TILDE	LATIN SMALL CONTRACTION PRO
76	LATIN SMALL LETTER P WITH BELT	LATIN SMALL CONTRACTION PRO WITH BELT
77	LATIN SMALL LETTER P WITH STROKE	LATIN SMALL CONTRACTION PER
78	LATIN SMALL LETTER Q WITH STROKE	LATIN SMALL CONTRACTION QUO
7D	LATIN SMALL LETTER Q SHARP S	LATIN CONTRACTION BUS
7E	LATIN SMALL LETTER MUSIC FLAT WITH HOOK	LATIN SMALL CONTRACTION IS

These characters represent contractions of Latin language prefixes, suffixes, and infixes most commonly found in manuscript material. ISO 5426-2 defined these characters for use with library data created in the United Kingdom at the British Library. It is believed that the BL is the largest library to have implemented this character set. This proposal suggests adding 18 of these characters (all but 42, 43, 44, and 48 above) as letterlike symbols within the ISO/IEC 10646 character code range 213B and 214C, justifying them as compatibility characters. The other four characters (42, 43, 44, and 48) should be added as combining diacritical marks within the ISO/IEC 10646 character code range 034F and 0352, also justifying them as compatibility characters. This would allow libraries such as the British Library, that have encoded bibliographic data using ISO 5425-2, to migrate to ISO/IEC 10646.

ISO 8957 - Hebrew

Of the 129 basic and extended Hebrew characters defined in ISO 8957, only eight characters cannot be mapped to equivalents in ISO/IEC 10646. In all cases these characters are used to encode marks found in Babylonian and Palestinian texts in the Hebrew script. They are as follows:

Hex	*Character Name*
42	HEBREW ACCENT ACUTE TSERE
43	HEBREW ACCENT GRAVE TSERE
57	HEBREW ACCENT BABYLONIAN QAMATS
5C	HEBREW ACCENT DAGESH
60	HEBREW ACCENT BABYLONIAN PATAH
61	HEBREW ACCENT BABYLONIAN QAMATS
62	HEBREW ACCENT BABYLONIAN DAGESH
66	HEBREW ACCENT ASTERISK

It is proposed that they be added in the ISO/IEC 10646 character code range 05C5 to 05CC. Addition of these characters to ISO/IEC 10646 is justified as compatibility characters for library data encoded using ISO 8957.

ISO 10754 - Extended Cyrillic for Non-Slavic Languages

Of the 93 graphic characters defined in ISO 10754, 22 cannot be mapped to equivalents in ISO/IEC 10646. The characters involved are as follows:

Hex	*Character Name*
24	COMBINING RIGHT DESCENDER
34	COMBINING LEFT DESCENDER
46	CYRILLIC SMALL LETTER KURDISH QA
47	CYRILLIC SMALL LETTER AISOR EL
49	CYRILLIC SMALL LETTER EL WITH MIDDLE HOOK HOOK
4A	CYRILLIC SMALL LETTER MORDVIN EL KA
4C	CYRILLIC SMALL LETTER CHUVASH NG
4E	CYRILLIC SMALL LETTER EN WITH MIDDLE HOOK HOOK
56	CYRILLIC CAPITAL LETTER KURDISH QA
57	CYRILLIC CAPITAL LETTER AISOR EL
59	CYRILLIC CAPITAL LETTER EL WITH MIDDLE HOOK MIDDLE HOOK
5A	CYRILLIC CAPITAL LETTER MORDVIN EL KA
5C	CYRILLIC CAPITAL LETTER CHUVASH NG
5E	CYRILLIC CAPITAL LETTER EN WITH MIDDLE HOOK MIDDLE HOOK
61	CYRILLIC SMALL LETTER SELKUP O IE
63	CYRILLIC SMALL LETTER ER KA
68	CYRILLIC SMALL LETTER KURDISH WE
6E	CYRILLIC SMALL LETTER YA IE
71	CYRILLIC CAPITAL LETTER SELKUP O IE
73	CYRILLIC CAPITAL LETTER ER KA
78	CYRILLIC CAPITAL LETTER KURDISH WE
7E	CYRILLIC CAPITAL LETTER YA IE

The first two characters are combining characters intended to be encoded following a base Cyrillic script character. The inclusion of a right or left descender on a base letter is a common technique in the Cyrillic script, used to differentiate between similar sounds. It was particularly popular to modify Cyrillic base letters with descenders for non-Slavic languages of Central Asia which adopted the Cyrillic alphabet during the era of the Soviet Union. JTC1 has rejected these characters on the grounds that most combinations of base letter a descender are already represented by precomposed characters in the Cyrillic repertoire. The use of a descender appears to be limited and not expanding for modern typography. This premise has not been thoroughly studied, but may be true. If any combinations of Cyrillic base letters with descenders are not covered by existing ISO/IEC 10646 precomposed characters, new characters could be added following the Character Encoding Model. TC46 is prepared to forego these combining marks.

The remaining 20 Cyrillic script characters that are not currently part of the ISO/IEC 10646 Extended Cyrillic repertoire need to be added to accommodate library use for existing Cyrillic script data. JTC1 has rejected some of these characters based on the grounds that they are ligatures that are sometimes represented as pairs of separate Cyrillic script letters in the basic Cyrillic repertoire. It is important to point out that the occasional representation of a ligatured character as two separate letters is not proof that two encoded characters should be used. Limitations of the typographical device used to create the text is usually why they are represented this way. Many Cyrillic script texts from the early to mid 20th century were type-written. Special characters and marks were often handwritten, or approximated due to the lack of special fonts or keyboards. Printed text, where leaded type could be produced, generally

provide the best representation of the alphabet for these languages. It is also important to point out that the Cyrillic script has a well-established tradition of typographical ligatures, particularly for sounds involving “L” and vowels. These letters coexist with isolated forms of the member consonants and vowels which can occur separately.

In a few cases, JTC1 has resisted adding characters from ISO 10754 based on similarity of the glyphs to characters in the Latin script repertoire. It is important to point out that many letters in the Cyrillic script are similar in shape to letters in the Latin script. This is not accidental. Cyrillic, Greek, and Latin (as well as many other scripts) have common ancestry. Letters from one script of the other have been borrowed and modified to meet the needs of different languages. It should not come as surprising that the Cyrillic script used for languages such as Kurdish should borrow from the Latin script for a letter to represent a gutteral sound. Since the glyphs “C” and “K” were already part of the Cyrillic script, it is perfectly logical that “Q” or something like it might be borrowed. It is important to note, using the Kurdish/Cyrillic “Q” as an example, that the actual glyph found in many texts is not the Latin letter Q. Even the capital Kurdish Q has often has the shape of a small “q”, only larger. Although the graphic representation is not the only consideration for uniqueness, the establishment of new Cyrillic script letters based on Latin or Greek script should not preclude additions to the Cyrillic repertoire. Their inclusion as distinct Cyrillic characters is particularly important for collation, where the character should *not* be treated as Latin script within an otherwise Cyrillic script string.

It is proposed that all 20 additional Cyrillic script letters from ISO 10754 be added to ISO/IEC 10646 as compatibility characters, without the requirement for exhaustive justification according to the Character Encoding Model. It should be pointed out that since this small repertoire of additional characters involves both small and capital letter forms, in reality it only adds 10 letters to the Cyrillic alphabet supported by ISO/IEC 10646.