

Universal Multiple-Octet Coded Character Set  
International Organization for Standardization  
Международная организация по стандартизации

Doc Type: Expert Contribution  
Title: Comments on the Mongolian Encoding Proposal, WG2 N1711  
Source: Ken Whistler  
Date: 1998-03-20

This document summarizes the comments and questions that I raised in the WG2 meeting regarding the proposed Mongolian encoding contained in WG2 N1711.

First of all, the great majority of the proposed characters have been stable and non-problematical for quite some time. In the proposal, I believe that the proposed characters shown at positions xx01..xx06, xx08..xx0C, xx10..xx19, xx20..xx77, xx80..xxA9 have no significant issues. There is consensus among the national bodies and experts regarding all of those characters. These characters and their encoded positions are completely unchanged from the preceding proposal (WG2 N1691), and the repertoire is unchanged from even earlier proposals.

The remaining problems concern xx01 MONGOLIAN SPACE, xx07 MONGOLIAN COMBINATION SYMBOL, and the 8 proposed format control characters at xx0D..xx0F and xx1B..xx1F.

## 1. xx00 MONGOLIAN SPACE

The MONGOLIAN SPACE is described in Appendix III, page 41, as a “non-breaking space”. It is unclear why this should be distinct or separately encoded from the already encoded character U+00A0 NO BREAK SPACE. The description shows various suffixes that are separated from the word stems in Mongolian by a visual space that does not constitute a word boundary. That function could be fulfilled by use of U+00A0, which would provide the visual spacing without inducing improper line-breaking behavior (if that is what is implied by word boundaries in Mongolian).

As regards searching operations in Mongolian, any operation which is looking for words would have to take the MONGOLIAN SPACE into account; any such operation could just as well be looking for and interpreting NO BREAK SPACE in the same way. Thus, there does not seem to be any strong reason to disunify MONGOLIAN SPACE from the NO BREAK SPACE.

As regards layout of the MONGOLIAN SPACE, while it is true that a NO BREAK SPACE in Latin text might have a different width than that expected for Mongolian text spacing, when a NO BREAK SPACE is laid out (vertically) in Mongolian text, it should be laid out according to the vertical layout logic for Mongolian, which could differ significantly. This is no different than having different glyphs and metrics for characters such as parentheses or corner bracket quotation marks when laid out horizontally and when laid out vertically. (Similar differences apply to the ordinary SPACE character.) The layout width is insufficient reason for separating the MONGOLIAN SPACE from the NO BREAK SPACE.

In my opinion, there must be further justification for the MONGOLIAN SPACE character before the case could be made for this disunification of the NO BREAK SPACE. This justification should show examples and demonstrate how use of U+00A0 NO BREAK SPACE for the functionality in Mongolian would produce results different than that required.

## 2. xx07 MONGOLIAN COMBINATION SYMBOL

The MONGOLIAN COMBINATION SYMBOL is shown as “?!”. While this might seem to be just a sequence of the already encoded characters “?” and “!”, the Inner Mongolian delegates have pointed out

that the “?!” punctuation is written as a side-by-side unit in vertical Mongolian text. To enable use of this mark in Mongolian plain text without having to engage in higher-level text formatting controls to embed a horizontal run in vertical text, it is reasonable to encode this element as a single character.

However, rather than treat this as Mongolian-specific punctuation, it would be more reasonable to simply encode this as a character in the General Punctuation block. My suggestion would be:

U+2047 QUESTION EXCLAMATION MARK

The reason for making this general punctuation is that it could also be used as a character in other vertically rendered scripts. It should not have a name specifically marking it as Mongolian.

### 3. xx1C..xx1F Mongolian positional format control characters

These positional format control characters (isolated, initial, medial, and final) are the part of the Mongolian proposal which has shown the most instability and indecision from document to document. They are intended to deal with the problem of being able to show Mongolian positional presentation forms in isolation, or in the middle of words under circumstances where the normal cursive joining rules have not applied.

The problem with this proposal is that the intended functionality of these additional characters is completely covered by two already-encoded characters:

U+200C ZERO WIDTH NON-JOINER

U+200D ZERO WIDTH JOINER

Those characters are encoded in ISO/IEC 10646 specifically to provide a mechanism for overriding the normal cursive joining rules in scripts such as Arabic (or Mongolian) which have cursive connections between characters and rules which determine when to display the isolated, initial, medial, or final form of a basic letter.

To clarify how the “joiner” and “non-joiner” characters can be used to accomplish what the proposed four Mongolian positional format control characters are intended for, I provide the following table, which exhaustively lists all the cases for which normal cursive joining apply and the override cases which employ the use of either joiner or non-joiner or both in combination.

Symbols used:

B Basic letter

O Isolated form

I Initial form

F Final form

M Medial form

\_ Space

J Joiner

NJ Non-joiner

In this table, in very abbreviated form, the desired display of the four positional forms of a basic letter is shown in the various possible contexts, under the “Display” column. Corresponding to each display, the backing store which is required to get this effect, with or without joiners or non-joiners, is shown. Thus, for example, “\_O\_” means show the isolated form of a letter between two spaces. The corresponding backing store is “\_ B \_”, which means space followed by the basic letter followed by space. To get a medial form shown between two spaces (e.g. “\_M\_”), the backing store should be “\_ J B J \_”, which means space followed by a joiner followed by the basic letter followed by a joiner followed by space. And so on.

Lower case letters are used to show other letters (and their positional forms) which may appear on either

side of the main letter (shown in upper case). Since the contextual letters either to the left or the right (or above and below, in the case of Mongolian), may themselves either join or not join to the letter they are next to, there are cases where a sequence of joiner plus non-joiner or non-joiner plus joiner may also be used to get the desired effect.

Display	Store	Display	Store
_O_	_ B _	_Fo_	_ J B NJ b _
_I_	_ B J _	_Mo_	_ J B J NJ b _
_F_	_ J B _	_iOf_	_ b J NJ B NJ J b _
_M_	_ J B J _	_iIf_	_ b J NJ B b _
_iO_	_ b J NJ B _	_iFf_	_ b B NJ J b _
_iI_	_ b J NJ B J _	_iMf_	_ b B b _
_iF_	_ b B _	_oOf_	_ b NJ B NJ J b _
_iM_	_ b B J _	_oIf_	_ b NJ J B b _
_oO_	_ b NJ B _	_oFf_	_ b NJ J B NJ J b _
_oI_	_ b NJ B J _	_oMf_	_ b NJ J B b _
_oF_	_ b NJ J B _	_iOo_	_ b J NJ B NJ b _
_oM_	_ b NJ J B J _	_iIo_	_ b J NJ B J NJ b _
_Of_	_ B NJ J b _	_iFo_	_ b B NJ b _
_If_	_ B b _	_iMo_	_ b B J NJ b _
_Ff_	_ J B NJ J b _	_oOo_	_ b NJ B NJ b _
_Mf_	_ J B b _	_oIo_	_ b NJ B J NJ b _
_Oo_	_ B NJ b _	_oFo_	_ b NJ B NJ b _
_Io_	_ B J NJ b _	_oMo_	_ b NJ J B J NJ b _

I believe that this table is logically complete, and covers all the combinations for which the Mongolian positional format control characters have been proposed. Therefore, those four characters are functionally mismatched duplicates of the joiner and non-joiner character; they should not be encoded in ISO/IEC 10646.

It should be noted that use of the already-encoded joiner and non-joiner characters would make it possible for system and application software developed for Arabic to be adapted fairly easily to work for Mongolian, since they would already have built in the required logic for handling joiners and non-joiners correctly.

In discussion of this analysis with the Inner Mongolian representatives from the Chinese delegation, there was some indication that there was still a requirement for a single positional indicator character, much like that shown in Document WG2 N 1691 at xx1C. If such a character is intended as a visible graphic character (a symbol) used in Mongolian educational texts to help explain how character conjoining works in the Mongolian script, it would be a useful addition to the Mongolian proposal. If, however, it is intended as a hidden control character, along the line of the analysis provided in WG2 N1691, then its use would be problematical, conflicting with the normal use of the joiner and non-joiner characters for controlling cursive connection in Mongolian.

#### 4. xx0D..xx0F Mongolian free variant selector characters

These three variant selector characters are proposed to cover the instances when the actual form that a character takes may take one of several variants, not predictable by position alone, but dependent on other factors (as, for example, gender) not derivable in any obvious algorithmic way. The Mongolian experts have verified that such variant forms do and must co-occur in ordinary Mongolian text, and the suggested solution of having several variant selector characters seems like a reasonable and parsimonious approach to this problem.

The issues with the free variant selector characters are twofold: 1. Are three of them actually needed? and 2. Should they be Mongolian-specific or coded for general use with other scripts which also require

explicit marking of free variant forms?

The question of whether three free variant selector characters are actually needed arises because the number actually used in the proposals has varied, even when addressing the same catalog of presentation forms shown in the Mongolian Reference Table (Appendix I). In the proposal in N1691, for example, two free variant selector characters are proposed, while in N1711, three are proposed. The extra one comes from a change in the analysis of presentation forms such as those for MONGOLIAN LETTER QA. In N1691, the Mongolian Reference Table shows two initial forms, two medial forms, two final forms, and two feminine isolate forms for QA. This requires the use of one free variant selector character to distinguish between each pair of positional forms. But in N1711, the Mongolian Reference Table shows exactly the same 8 presentation forms, but changes the characterization of the two final forms to be “second medial forms”. This results in 4 different “medial” forms, necessitating the use of 3 free variant selector characters to distinguish them all. I believe that removing the 4 Mongolian positional format control characters and using the joiner and non-joiner characters instead enables a return to an analysis more like that presented in N1691, in which only two free variant selector characters are actually needed to make all the required distinctions for most characters. However, at least one example still appears to exist (MONGOLIAN LETTER MANCHU I, number 115 of the Basic Characters in the Mongolian Reference Table) that is best treated with three variant selector forms.

Granted that three free variant selector characters are needed for Mongolian, the next question is whether these should be specific to the Mongolian script or should be part of a set of generic free variant selector characters that can be used with any script. This question should be addressed by WG2 and the Unicode Technical Committee when considering where to encode the three free variant selector characters for Mongolian. (Such variant selection issues have already surfaced for Tibetan and for CJK characters, and may also be an issue for many historic scripts.)

## **5. xx1B MONGOLIAN VOWEL SEPARATOR**

This character is proposed to handle a particular Mongolian presentation case, where a consonant followed by the vowel a or e is shown in a special way, with the consonant and vowel not joined, the a or e in a variant final form, and in some cases with dots from the consonant rendered inside the loop of the variant final form for the vowel.

Technically, a sequence such as ML. NA + MVS + ML. A could equally well be expressed by the sequence ML. NA + non-joiner + ML.A + FVS2. In other words, the non-junction of the consonant and vowel is indicated by the non-joiner character, and the variant form of the final vowel is selected by use of the free variant selector character number 2. This would be sufficient to make the required distinction in plain text and could be used for text interchange.

However, the Inner Mongolian representatives from the Chinese delegation have indicated an implementation preference for having a single character for this function. I believe that the addition of such a character would be less desirable, but not as problematical as the positional format control characters. If China, Mongolia, and the Mongolian experts insist on its inclusion, it would probably not cause problems for other text encoding. However, I would suggest that the name be changed to:

### **xx1C MONGOLIAN VOWEL ZERO WIDTH NON-JOINER**

This character would be functioning as another zero-width non-joiner, but would only have special formatting behavior when occurring in Mongolian text.

I would further suggest that review of this character take into account the relative benefits versus costs of this disunification of the existing ZERO WIDTH NON-JOINER character to create this special Mongolian character. (There are clearly benefits for Mongolian, but one of the drawbacks is that existing applications which use non-joiners would now have to check for two possible values for non-joiners instead of just one.)

## **6. xx08 MONGOLIAN TODO SOFT HYPHEN**

This character seems clearly justified. However, its name should be changed to MONGOLIAN TODO HYPHEN, unless there is evidence that it behaves like a soft hyphen. Clear examples of usage should be provided for discussion. (For example, it is rendered at the start of the continuation line, rather than at the end of a word which is broken across a line.) The identity of this as a soft hyphen versus a regular hyphen depends on whether this character is rendered internal to a line or is hidden in a line when coded in the middle of a word.

## **7. General**

It would be very helpful in the process of speeding the Mongolian proposal towards a swift and successful balloting for the Chinese delegation to provide a short but detailed document which addresses the questions and observations I have made here. It would be very important to include examples along with the explanations, and in particular to focus on making detailed justifications for any disunifications.

As I have pointed out, the proposed positional format control characters and the MONGOLIAN VOWEL SEPARATOR represent disunifications of the already existing non-joiner and/or joiner characters. And the proposed MONGOLIAN SPACE character represents a disunification of the already existing NO BREAK SPACE. These are the disunifications which are causing most of the objections to the proposal. Those objections should be addressed in detail, or the proposal should drop those characters.