L2/23-065

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 to encode a blank character for Khitan Small Script**
Source: **Andrew West**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2023-03-01**

1. Introduction

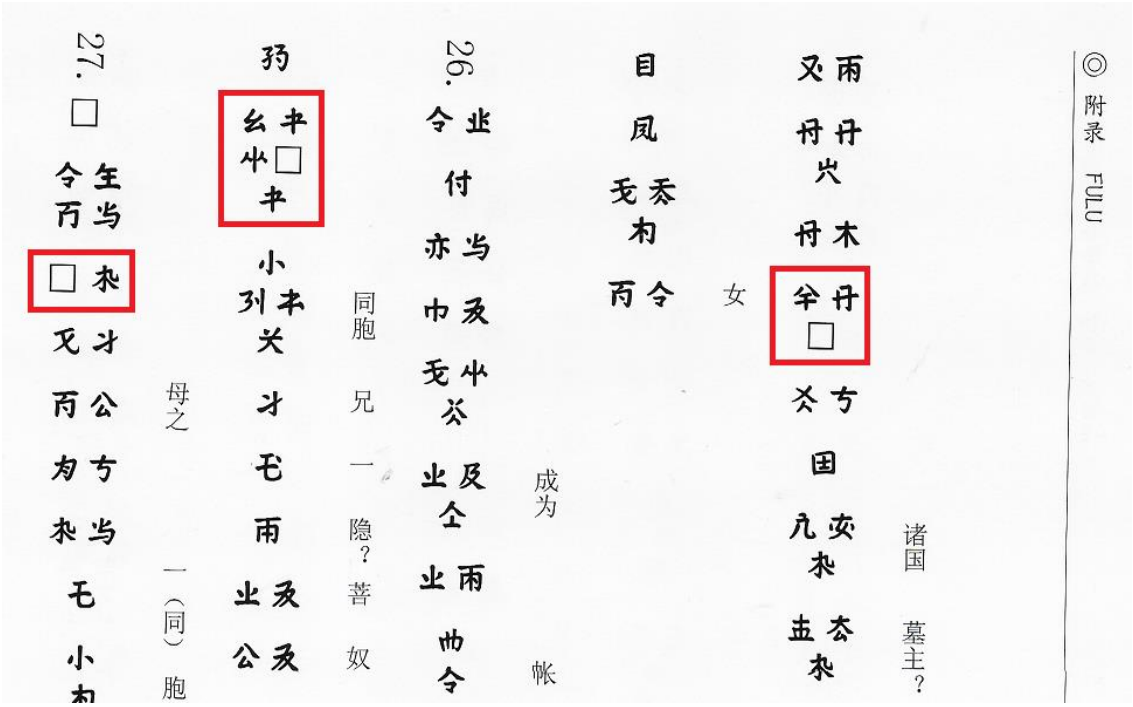
This document proposes the addition of a blank character representing a lost or illegible character for the Khitan Small Script. As the Khitan Small Script is an historic script, its primary use case is for academic transcriptions of Khitan text, which largely survives in the form of epitaphs engraved on stone tablets. Often the text on such stone tablets is damaged or unclear, meaning that some Khitan characters cannot be read. The standard convention in such cases is to represent the illegible or lost character with a hollow box with solid or dotted edges. Where the lost character is a single, isolated logogram, then we could use U+25A1 WHITE SQUARE □ (as is the case for Chinese, Tangut, etc.), but where the lost character is part of a combined cluster of characters, the blank glyph needs to participate in Khitan Small Script shaping behaviour, and so a script-specific character is preferred.

The need for a specific character to represent a missing or illegible character in a cluster was not considered with sufficient care when the original proposal to encode the Khitan Small Script was made, and we erroneously assumed that U+25A1 or U+2B1A could be used (see L2/16-113R / WG2 N4725R p. 116). It is only now that actual Khitan Small Script texts are being digitized and made available on the internet that we have realized that U+25A1 and U+2B1A are not suitable. In particular, the author of this document has completed a Unicode transcription of the long Hudujin Shenmi inscription (1,528 logographs and clusters, in total 3,828 Unicode characters) with the intention to upload it to Wikisource (https://wikisource.org/wiki/Main_Page/Khitan), but as the text includes thirty-eight clusters with one or more missing characters which will not render correctly, the digital text cannot be distributed in its current form. The need for the proposed character is therefore rather urgent.

2. Examples

In Fig. 1, the author uses a holllow box with solid edges to represent the missing or illegible character. As the computer technology used to lay out clusters in this example is quite primitive (it is not done at the font level), and simply positions individual characters in a block, without modifying their height or width to reflect the overall shape of the cluster, all characters are the same size, and so the hollow boxes are also all the same size.

Fig. 1. Khitan Small Script clusters with hollow boxes

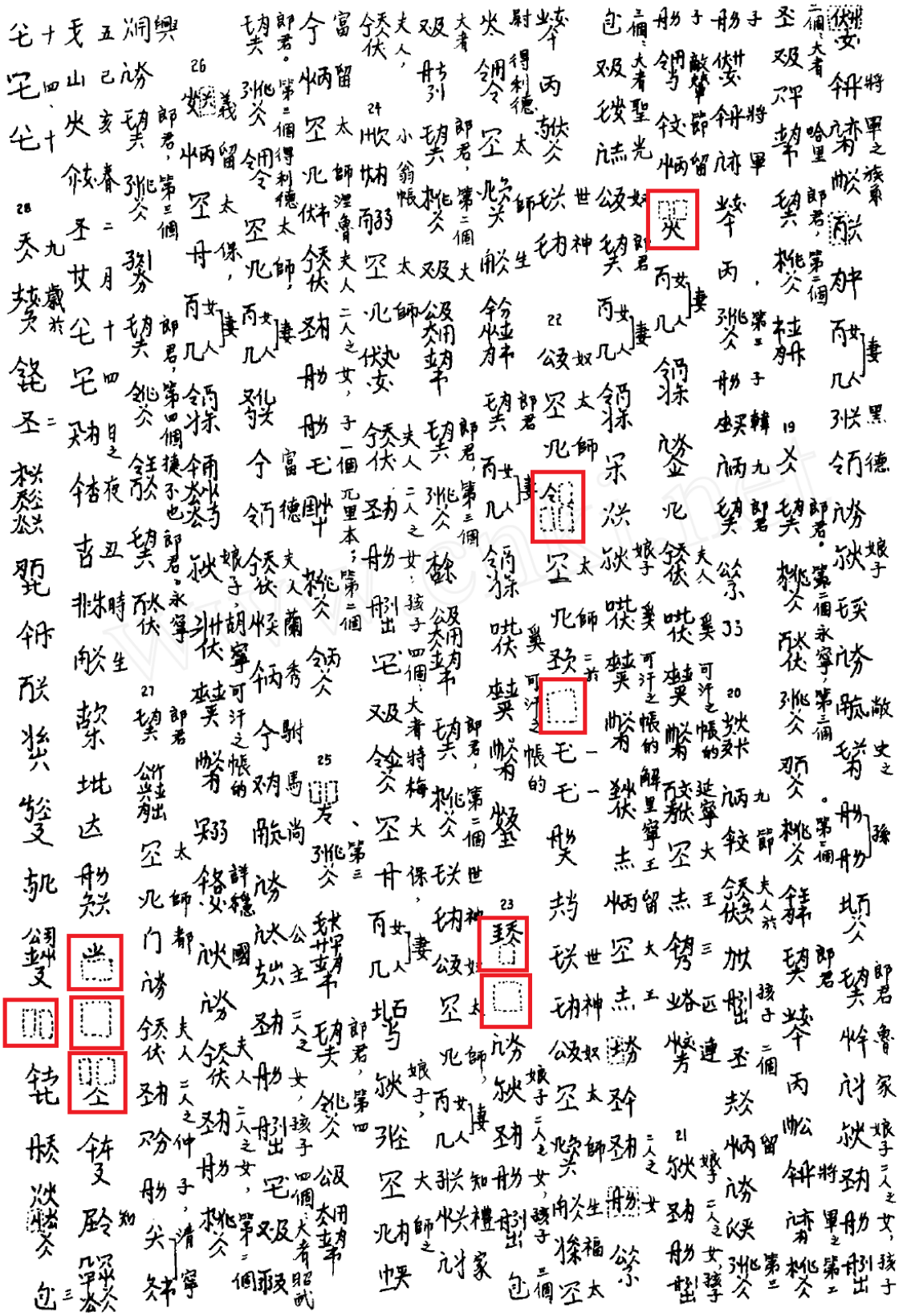


Wú Yīngzhé 吴英喆: 契丹小字《胡睹堇审密墓志铭》考释 [Study of the Khitan Small Script *Epitaph for Hudujin Shenmi*] (2011) p. 251

In Fig. 2, the author gives a hand-written transcription of a Khitan Small Script epitaph inscription, with Chinese glosses added to the side. This author represents missing or illegible characters with a hollow box with dotted edges. As characters which are placed side-by-side in a cluster are narrower than characters which occur in isolation, the size of the dotted box varies: a dotted box representing a single character in isolation is square; whereas a dotted box representing a lateral character in a cluster is rectangular (taller than wider); and a dotted box representing an initial or terminal character in a cluster is rectangular (wider than taller). This modification of the width and height of characters inside a cluster better reflects actual Khitan Small Script orthography than Fig. 1.

Note that in Fig. 2, the author draws a dotted box around a character inside a cluster for which the reading is uncertain. I am not proposing to encode a script-specific combining dotted box for this usage.

Fig. 2. Khitan Small Script clusters with dotted boxes



Zhèng Xiǎoguāng 郑晓光, 契丹小字《耶律永宁郎君墓志铭》考释 [Study of the Khitan Small Script *Epitaph for Yelü Yongning Langjun*] (*Minzu Yuwen* 民族语文 2002.2) p. 67

3. Discussion

The situation for Khitan Small Script is analogous to that of Egyptian Hieroglyphs, where there is a requirement to represent lost or damaged hieroglyphs in transcribed text, but due to the complex structural arrangement of Egyptian hieroglyphs, it has been necessary to encode U+13443 EGYPTIAN HIEROGLYPH LOST SIGN and various other characters (see Fig. 3).

Fig. 3. The Unicode Standard v. 15.0 p. 453

Lost Signs. To indicate text that had existed earlier, but was later destroyed, U+13443 EGYPTIAN HIEROGLYPH LOST SIGN, U+13444 EGYPTIAN HIEROGLYPH HALF LOST SIGN, U+13445 EGYPTIAN HIEROGLYPH TALL LOST SIGN and U+13446 EGYPTIAN HIEROGLYPH WIDE LOST SIGN are used. Some of these lost signs are shown in *Figure 11-8* next to other extant signs. The "lost signs" may appear in groups with other signs and are generally rendered as shaded squares or rectangles with whitespace between the signs. If continuous shading is required without whitespace between the signs, then U+FE00 VARIATION SELECTOR-1 immediately follows the blank lost character, so that no whitespace appears.

Figure 11-8. Use of Lost Signs

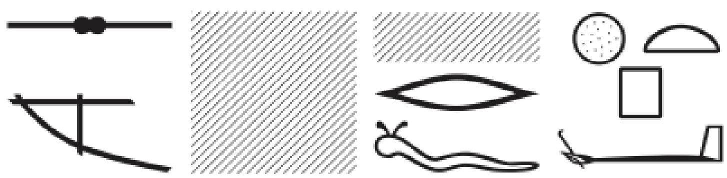


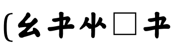
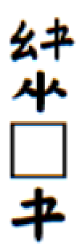
Fig. 4 shows my attempts to represent the middle cluster shown in Fig. 1 ( in linear format) using two experimental fonts with complex shaping behaviour. In the first font I used U+25A1 with positional glyph variants for the blank glyph, but under Windows 10 the OpenType feature used for shaping Khitan text was not applied to U+25A1, and so the cluster shaping was broken. In the second font I used the reserved code point U+18CFF at the end of the Khitan Small Script block as a blank character with positional glyph variants (i.e. □ in isolation, but with varying width and height when part of a cluster), which produced the expected shaping behaviour.

Fig. 4. Experimental Khitan Small Script fonts

A. Font using U+25A1



B. Font using U+18CFF

In conclusion, I believe that it is necessary and appropriate to encode a single blank character in the Khitan Small Script block to enable representation of lost or illegible characters within complex clusters of Khitan small script characters.

4. Properties

Recommended code point: U+18CFF

Character name: KHITAN SMALL SCRIPT CHARACTER-18CFF

Code chart annotation: represents a lost or illegible character

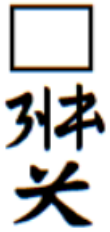
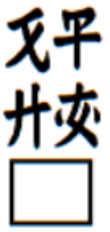
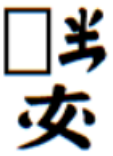
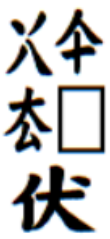
Reference glyph: □

All other Unicode properties should be the same as for all other encoded characters in the Khitan Small Script block.

5. Font Implementation

The recommended font implementation would be to provide a square hollow box with solid, dashed, or dotted edges as the base glyph for the new character. For fonts that implement clustering behaviour, rectangular variants of this character should be applied as appropriate. Examples from my test font are shown below.

Table 1. Positional glyph variants for the blank character

Isolate (base glyph)	Cluster Initial	Cluster Terminal	Cluster Lateral	Cluster Lateral (Medial)	Cluster Lateral (Medial)
					

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** ***Proposal to encode a blank character for Khitan Small Script***
2. Requester's name: *Andrew West*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *1 March 2023*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *Yes*
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters):
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Yes*
      Name of the existing block: *Khitan Small Script*
2. Number of characters in proposal: *1*
3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary	X	B.1-Specialized (small collection)		B.2-Specialized (large collection)	
C-Major extinct		D-Attested extinct		E-Minor extinct	
F-Archaic Hieroglyphic or Ideographic		G-Obscure or questionable usage symbols			

4. Is a repertoire including character names provided? *Yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *N/A*
   b. Are the character shapes attached in a legible form suitable for review? *Yes*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   Andrew West
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   Andrew West
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *No*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *Yes*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *Yes*

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *Yes*
   If YES, with whom? *Khitan small script experts*
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *No*
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *Common*
   Reference:
5. Are the proposed characters in current use by the user community? *Yes*
   If YES, where? Reference:
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *No*
   If YES, is a rationale provided?
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *N/A*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *Yes*
   If YES, is a rationale for its inclusion provided? *Yes*
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? *Yes*
    If YES, is a rationale for its inclusion provided? *Yes*
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? *No*
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *No*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility characters? *No*
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: