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

Doc Type: **Working Group Document**
Title: **Technical update on Proposal to encode Jurchen Small Script characters**
Source: **Viacheslav Zaytsev and Andrew West**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2025-06-10**

1. Introduction

In our previous document, "Proposal to encode Jurchen Small Script characters" (WG2 N5309 = L2/25-152) (2025-05-22), we provided a comprehensive academic study addressing the historical, philological, and technical considerations for encoding characters of the Jurchen Small Script (JSS), a poorly preserved script attested on a single short inscription separately engraved on three gold and silver *páizǐ* 牌子 (travel passes or symbols of authority) unearthed in northeast China during the 1970s and 1980s. The proposal included detailed historical background, evidence for the script's identity, and an evaluation of encoding options, recommending encoding within the Khitan Small Script (KSS) block (U+18B00–U+18CFF) due to structural similarities and limited evidence.

Following feedback from the Script Encoding Working Group (SEWG) on June 6, 2025, which endorsed Option 1 (encoding Jurchen Small Script characters within the Khitan Small Script block), we present this technical update to formalize the code points, character names, Unicode character properties, and annotations for the proposed characters, including subsection titles and cross-references for the code charts and names list, as requested. This document also provides a draft description for the Unicode Core Specification and an updated ISO/IEC JTC1/SC2/WG2 proposal form to support the formal submission. Readers seeking detailed historical and philological analysis are referred to WG2 N5309 (L2/25-152), as this update focuses solely on the technical requirements for encoding.

2. Proposed Characters and Code Point Allocation

Based on the SEWG's endorsement of Option 1, we propose encoding five new Jurchen Small Script characters in the Khitan Small Script block (U+18B00–U+18CFF) at code points U+18CD6 through U+18CDA. This allocation leaves 36 free code points (U+18CDB–U+18CFE) in the Khitan Small Script block for future additions. The proposed characters, their code points, and their glyph representations are presented in Table 1.

Table 1: Proposed additions to the Khitan Small Script block

Code Point	Glyph	Character Name
U+18CD6	𘳖	KHITAN SMALL SCRIPT CHARACTER-18CD6
U+18CD7	𘳗	KHITAN SMALL SCRIPT CHARACTER-18CD7
U+18CD8	𘳘	KHITAN SMALL SCRIPT CHARACTER-18CD8
U+18CD9	𘳙	KHITAN SMALL SCRIPT CHARACTER-18CD9
U+18CDA	𘳚	KHITAN SMALL SCRIPT CHARACTER-18CDA

Additionally, we propose unifying the Jurchen Small Script character (𘳛) with the existing Khitan Small Script character at U+18C3E (𘰾), with a glyph modification to reflect the angled legs characteristic of the Jurchen glyph, as outlined in WG2 N5309 (Section 2, pp. 3–5). The proposed glyph modification is presented in Table 2.

Table 2: Proposed glyph modification for existing Khitan Small Script character

Code Point	Current Glyph	New Glyph	Character Name
U+18C3E	𘰾	𘰾	KHITAN SMALL SCRIPT CHARACTER-18C3E

3. Rationale for Character Naming

The proposed characters at U+18CD6–U+18CDA and the unified character at U+18C3E are named KHITAN SMALL SCRIPT CHARACTER-XXXXX, aligning with the naming convention of the Khitan Small Script block (U+18B00–U+18CFF), rather than using JURCHEN SMALL SCRIPT CHARACTER-XXXXX (see Table 1 and Table 2). This decision addresses concerns about algorithmic naming consistency and ensures compatibility with existing software and rendering systems. The rationale is as follows:

Algorithmic Consistency: Using a uniform prefix (KHITAN SMALL SCRIPT CHARACTER-) for all characters in the KSS block simplifies parsing and processing by software, font designers, and text-processing tools, which rely on predictable naming patterns within a single block. A mixed naming scheme with JURCHEN SMALL SCRIPT CHARACTER- prefixes could complicate implementation, as it deviates from the established KSS block convention.

Unification and Limited Evidence: Unification of JSS and KSS (encoding JSS characters within the KSS block), as proposed in WG2 N5309 and endorsed by SEWG, is based on their structural similarities, such as character clustering, and the limited JSS corpus (six characters from *páizǐ* inscriptions). As it is unclear whether JSS was entirely distinct from KSS or borrowed some KSS characters unchanged, a consistent naming convention supports the provisional unification, avoiding premature differentiation that could hinder the ability to build clusters of mixed KSS and JSS characters, which is potentially important given the uncertainty about their relationship.

Preserving JSS Identity: To address concerns about obscuring JSS's historical and cultural identity, annotations in ***NamesList.txt*** (see Section 4.2) and the Unicode Core Specification (see Section 5) explicitly indicate the JSS usage of these characters. Unlike the Mongolian block, where Manchu characters are named MONGOLIAN LETTER MANCHU ... to indicate language-specific usage, JSS characters are named generically as KHITAN SMALL SCRIPT CHARACTER-XXXXX due to the tentative identification and small corpus, with annotations providing specificity. This ensures scholarly clarity without disrupting block uniformity.

Rendering and Implementation Simplicity: Consistent naming supports seamless integration with the KSS rendering system, which handles character clusters. A distinct naming convention could imply a separate script identity, potentially requiring unnecessary adjustments to font design or rendering logic, despite JSS's compatibility with KSS clustering behavior.

This naming approach aligns with the SEWG's unification decision, leverages the limited evidence for JSS, and maintains flexibility for future reclassification if additional JSS inscriptions are discovered, as noted in WG2 N5309 (Section 5, pp. 38–39).

4. Unicode Character Database (UCD) Entries

The Unicode Character Database (UCD) entries for the proposed characters and the modified character at U+18C3E are provided below, following the format specified in the Unicode Standard.

4.1. Unicode Character Properties (UnicodeData.txt)

```
18CD6;KHITAN SMALL SCRIPT CHARACTER-18CD6;Lo;0;L;;;;N;;;;;
18CD7;KHITAN SMALL SCRIPT CHARACTER-18CD7;Lo;0;L;;;;N;;;;;
18CD8;KHITAN SMALL SCRIPT CHARACTER-18CD8;Lo;0;L;;;;N;;;;;
18CD9;KHITAN SMALL SCRIPT CHARACTER-18CD9;Lo;0;L;;;;N;;;;;
18CDA;KHITAN SMALL SCRIPT CHARACTER-18CDA;Lo;0;L;;;;N;;;;;
```

No property change:

```
18C3E;KHITAN SMALL SCRIPT CHARACTER-18C3E;Lo;0;L;;;;N;;;;;
```

4.2. Annotations, Subsections, and Cross-References for Code Charts and Names List (NamesList.txt)

```
18C3E  KHITAN SMALL SCRIPT CHARACTER-18C3E
       * used in Jurchen Small Script and possibly in Khitan Small Script
       * glyph reflects the Jurchen Small Script form with angled legs

<…>

@             Jurchen Small Script characters
@+            Characters tentatively identified as Jurchen Small Script, encoded in
              the Khitan Small Script block due to similar clustering structure and
              limited evidence. This set also includes 18C3E.
              x (khitan small script character-18c3e - 18C3E)
18CD6  KHITAN SMALL SCRIPT CHARACTER-18CD6
       * used in Jurchen Small Script
18CD7  KHITAN SMALL SCRIPT CHARACTER-18CD7
       * used in Jurchen Small Script
18CD8  KHITAN SMALL SCRIPT CHARACTER-18CD8
       * used in Jurchen Small Script
18CD9  KHITAN SMALL SCRIPT CHARACTER-18CD9
       * used in Jurchen Small Script
18CDA  KHITAN SMALL SCRIPT CHARACTER-18CDA
       * used in Jurchen Small Script
```

The proposed ***NamesList.txt*** entries above are suggested to guide the formulation of final annotations, using “x” to indicate cross-references to other characters and “*” for comments on usage or glyph details, per Unicode conventions. These are offered as a draft for consideration, and refinements are welcome.

The identity of U+18CD6–U+18CDA as JSS characters is indicated through a subsection header and per-character annotations (“* used in Jurchen Small Script”), with a similar annotation applied to U+18C3E. These annotations indicate the Jurchen identity of the characters, as justified in the Naming Rationale (Section 3), since a subsection title alone may not suffice for individual character queries.

The entry for U+18C3E also conveys that its glyph was modified to reflect the Jurchen Small Script form with angled legs and its unification with the originally encoded Khitan Small Script character. Its use in KSS is attested solely in modern sources, suggesting it may be a "ghost character" from a Khitan textual perspective, while the Jurchen form is confirmed by primary sources (see WG2 N5309, Section 2, pp. 3–5 for details).

4.3. Script Property (Scripts.txt)

We suggest assigning the script property value **Khitan_Small_Script** to the proposed characters (U+18CD6–U+18CDA) and preserving it for the modified character at U+18C3E in ***Scripts.txt***, as shown below. Rationale for this suggestion follows after.

Updated ***Scripts.txt*** Entries:

```
# ================================================

16FE4          ; Khitan_Small_Script # Mn       KHITAN SMALL SCRIPT FILLER
18B00..18CDA   ; Khitan_Small_Script # Lo [475] KHITAN SMALL SCRIPT CHARACTER-18B00..
                                                KHITAN SMALL SCRIPT CHARACTER-18CDA
18CFF          ; Khitan_Small_Script # Lo       KHITAN SMALL SCRIPT CHARACTER-18CFF

# Total code points: 477

# ================================================
```

Existing rendering systems for KSS, which support clustering of characters, are fully compatible with JSS, ensuring consistency in text processing and font design. The assignment of the **Khitan_Small_Script** script property is critical for the following reasons:

> **Support for Mixed KSS/JSS Clusters**: As it is unknown whether JSS was entirely distinct from KSS or borrowed some KSS characters unchanged, it is potentially important to be able to build clusters of mixed KSS and JSS characters. Assigning the same **Khitan_Small_Script** script property ensures that rendering engines treat these characters as part of the same script, preventing text segmentation at script boundaries that could break cluster formation or affect word selection.
>
> **Compatibility with Blank and Format Characters:** The KSS blank character at U+18CFF, used to represent obscured or missing characters in clusters, and the KSS format character at U+16FE4, used to form Type B clusters, have the **Khitan_Small_Script** script property. Assigning the same property to JSS characters enables their use with U+18CFF and U+16FE4 in JSS text, maintaining rendering consistency. A distinct "Jurchen Small Script" property would preclude this compatibility, requiring new blank or format characters or property updates, which is impractical given the limited JSS corpus.

Precedent of Unified Scripts: The assignment follows the precedent of the Mongolian block, where Manchu, Todo, and Sibe characters are assigned the **Mongolian** script property in ***Scripts.txt*** (e.g., for characters named MONGOLIAN LETTER MANCHU ...), despite their distinct identities. Similarly, JSS characters, including the modified U+18C3E, are assigned the **Khitan_Small_Script** property, reflecting their unification with KSS due to structural similarities and tentative identification.

Avoiding Premature Differentiation: Given the limited JSS evidence (six characters) and uncertainty about its distinctness from KSS, a new ISO 15924 code for "Jurchen Small Script" is premature. The **Khitan_Small_Script** property supports the provisional unification endorsed by the SEWG, with JSS's identity preserved through ***NamesList.txt*** annotations (see Section 4.2) and Core Specification text (see Section 5).

This script property assignment complements the character naming approach (see Section 3) by ensuring rendering compatibility with KSS clustering and format characters, while allowing for potential reclassification should further JSS evidence emerge, as noted in WG2 N5309 (Section 5, pp. 38–39).

5. Description for the Unicode Core Specification

The following draft description is proposed for inclusion in the Unicode Core Specification to summarize the encoding of Jurchen Small Script within the Khitan Small Script block. Additions or corrections to the current version of the Core Specification are highlighted in yellow.

18.12 Khitan Small Script

18.12.1 Khitan Small Script: U+18B00–U+18CFF

The Khitan Small Script block encodes characters for two scripts: the Khitan Small Script, used by the Khitan people to write the Khitan language, and the Jurchen Small Script, a script historically documented as used by the Jurchen people to write the Jurchen language, though surviving inscriptions are only tentatively identified as this script. Given the uncertainty of this identification and the apparent structural similarities of these inscriptions to Khitan Small Script, particularly in character clustering, Jurchen Small Script characters are encoded within this block due to the limited evidence available.

Khitan Small Script was one of two scripts used by the Khitan people of Northern China to write the Khitan language during the Liao dynasty (907–1125 CE), the Qara Khitai empire (or Western Liao dynasty, 1124–1218), and the Jin dynasty (1115–1234). The other script is known as Khitan Large Script. Both scripts are only partially deciphered today but were used over the same time period, in the same geographical area, and for the same functions.

Khitan Small Script was created about 925 by Yelü Diela, and its creation is said to have been inspired by the Uyghur script, although there appear to be few similarities between the two scripts. The main source of texts in Khitan Small Script are funerary epitaphs engraved on stone tablets and buried with members of Khitan royalty and aristocracy. The script also appears on walls and monuments, as well as on bronze mirrors, tallies, non-circulation coins, and a single jade cup.

Jurchen Small Script, tentatively identified in a small corpus of inscriptions, is attributed to the Jurchen, a Tungusic people in Northern China who founded the Jin dynasty (1115–1234) under Emperor Taizu (Wanyan Aguda, r. 1115–1123). The script was promulgated in 1138 under Emperor Xizong (r. 1135–1150) and used concurrently with the Jurchen Large Script, created in 1119 by Wanyan Xiyin and possibly Yelu. Evidence for Jurchen Small Script is confined to a single inscription of six characters in two clusters on gold and silver *páizǐ* (travel passes or symbols of authority) unearthed in northeast China during the 1970s and 1980s. Due to this scant corpus, no definitive statements can be made about its structure, use, or decipherment, and the identification of these inscriptions as Jurchen Small Script remains unconfirmed. Current understanding assumes it mimics the Khitan Small

Script, described below, as the *páizǐ* inscriptions exhibit a cluster structure resembling Khitan Small Script.

Unification of Characters in Khitan Small Script Block. The Khitan Small Script block unifies characters for Khitan Small Script and Jurchen Small Script. Khitan Small Script characters occupy U+18B00–U+18CD5, while Jurchen Small Script characters are encoded at U+18CD6–U+18CDA, with one character at U+18C3E used for both scripts. This unification is primarily driven by their assumed kinship, as evidenced by shared structural similarities such as character clustering, and by the limited Jurchen Small Script corpus, which practically precludes a separate block. Both scripts also utilize a blank character at U+18CFF and a format character at U+16FE4 (see below).

<…>

The following subsection (***Structure***) remains unchanged from the current Unicode Core Specification.

<…>

Character Names. The Khitan Small Script and Jurchen Small Script characters are named sequentially by prefixing "KHITAN SMALL SCRIPT CHARACTER-" to the code point, with the exception of one format control character, U+16FE4 KHITAN SMALL SCRIPT FILLER. The filler character is located in the Ideographic Symbols and Punctuation block. The naming of Jurchen Small Script characters with the Khitan prefix reflects their unification within the Khitan Small Script block. This approach primarily addresses concerns about algorithmic consistency and compatibility with existing systems. It also supports the provisional unification due to limited evidence and the uncertainty of the relationship between both scripts, allowing for flexible future reclassification. For user convenience, Jurchen characters are also provided with annotations that reflect their Jurchen identity.

<…>

The following subsections (***Phonogram Clusters***, ***Iteration Mark***, ***Obscured or Missing Characters***, and ***Figure 18-17***) remain unchanged from the current Unicode Core Specification.

6. Recommendations and Notes for Implementation

Font Design: The proposed glyphs (𘓐, 𘓑, 𘓒, 𘓓, 𘓔, 𘓕) should use the same font style as the existing Khitan Small Script code chart font, as noted in WG2 N5309. The glyph at U+18C3E should be updated to reflect the angled legs of the Jurchen form (𘓕).

Rendering Support: No changes to the existing Khitan Small Script rendering system are required, other than to extend the range of Khitan Small Script characters to include U+18CD6 through U+18CDA, as the Jurchen Small Script characters follow the same clustering behavior.

Future Considerations: If additional Jurchen Small Script inscriptions are discovered, a dedicated block may be proposed, as discussed in WG2 N5309 (Section 5, pp. 38–39).

Recommendation: The block name and character names for the proposed “Jurchen” script (see WG2 N5261R), covering the presumed Jurchen Large Script, risk ambiguity as “Jurchen” is a generic term. Naming the block “Jurchen Large Script” and using the algorithmic character names “JURCHEN LARGE SCRIPT CHARACTER-XXXXX” is advisable to distinguish it from the Jurchen Small Script. This scheme follows the model used for Khitan Large and Small Scripts.

7. Conclusion

This technical update addresses the Script Encoding Working Group (SEWG)'s feedback from June 6, 2025. It formalizes the code point allocation for five Jurchen Small Script characters within the Khitan Small Script block (U+18CD6–U+18CDA) and unifies one character with U+18C3E (with a glyph modification). Additionally, it provides Unicode character properties, code chart and names list annotations (including subsections and cross-references), and a script property recommendation. In response to potential concerns about naming and script property consistency, the characters are named "KHITAN SMALL SCRIPT CHARACTER-XXXXX" and assigned the **Khitan_Small_Script** script property, ensuring rendering compatibility, including support for mixed KSS/JSS clusters and use of the U+18CFF blank character and U+16FE4 format character, while annotations indicate the Jurchen Small Script identity of these characters. A draft description for the Unicode Core Specification is provided to ensure clarity in the standard. The attached ISO/IEC JTC1/SC2/WG2 proposal form, updated to reflect recent revisions, supports the formal submission.

We await further feedback from the Unicode Technical Committee (UTC) and ISO/IEC JTC1/SC2/WG2. Should additional recommendations arise, only this technical update will be revised, as the foundational research in WG2 N5309 (L2/25-152) is complete and serves as the definitive reference for the historical, philological, and encoding aspects of Jurchen Small Script.

8. References

N5309. *Vicheslav Zaytsev, Andrew West*. Proposal to encode Jurchen Small Script characters. ISO/IEC JTC1/SC2/WG2 N5309 = L2/25-152. 2025-05-22.

URL: https://www.unicode.org/wg2/docs/n5309-Proposal-JurchenSmallScript.pdf
URL: https://www.unicode.org/L2/L2025/25152-wg2-n5309.pdf

N5261R. *Andrew West* (魏安), *Sun Bojun* (孙伯君), *Zhōnghuá Zìkù* 中华字库 *project*. Proposal to Encode the Jurchen Script. JTC1/SC2/WG2 N5261R = L2/24-139. 2024-06-12.

URL: https://www.unicode.org/wg2/docs/n5261R-jurchen-ideographs.pdf
URL: https://www.unicode.org/L2/L2024/24139-n5261r-jurchen-ideographs.pdf

Proposal Summary Form

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://www.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://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** *Technical update on Proposal to encode Jurchen Small Script characters*
2. Requester's name: *Viacheslav Zaytsev and Andrew West*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *2025-06-10*
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): *NO*
      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: *5*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary ___ B.1-Specialized (small collection) ___ B.2-Specialized (large collection) ___
   C-Major extinct ___ D-Attested extinct ___ E-Minor extinct ___
   F-Archaic Hieroglyphic or Ideographic X G-Obscure or questionable usage symbols ___
4. Is a repertoire including character names provided? *N/A*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?
   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? *YES*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *NO*
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)? *NO*

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: N4102-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? *Script Encoding Working Group*
   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) *rare*
   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)? *YES*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *NO*
   If YES, is a rationale for its inclusion provided?
   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: *See WG2 N5309 Section 2*
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: