

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 nine Tangut ideographs and six Tangut components

Source: Andrew West, Viacheslav Zaytsev (Institute of Oriental Manuscripts, Russian Academy of Sciences), Jia Changye (Ningxia Academy of Social Sciences), Jing Yongshi (Beifang University of Nationalities), Sun Bojun (Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences)

Status: Individual Contribution

Action: For consideration by JTC1/SC2/WG2 and UTC

Date: 2019-05-27

1. Introduction

At the Ad hoc meeting on Tangut held at Yinchuan, China in August 2016, under the auspices of the Script Encoding Initiative, Professors Jia Changye and Jing Yongshi reported that they had identified a number of misunified Tangut ideographs (see WG2 N4736; L2/16-243). The ideographs in question each have two unrelated meanings with separate entries in Li Fanwen's 2008 *Tangut-Chinese Dictionary* (*Xià-Hàn zìdiǎn* 夏漢字典), but because they have identical glyphs in Li Fanwen's dictionary and all other modern sources each of the two meanings were unified as a single encoded character. The recent research by Jia and Jing indicates that there are subtle but systematic glyph differences that distinguish the two readings and meanings of each of these encoded characters, as listed in their *Xixia zìfú jí shǔxìng biāozhù biǎo (cǎogǎo)* 西夏字符及属性标注表(草稿) [Table of Xixia Characters with Annotated Properties (Draft)] (August 2016). Their document also identifies five components which should each be disunified into two encoded characters.

As these glyph differences had not been previously noted by Tangutologists, Andrew West and Viacheslav Zaytsev were tasked with investigating the issue and suggesting a solution. West and Zaytsev's investigation (WG2 N5031; L2/19-064) provides detailed evidence supporting the systematic glyph differences noted by Jia and Jing, and recommends disunifying the relevant Tangut ideographs and Tangut components.

On the basis of the research by Jia and Jing, and the evidence provided by West and Zaytsev in WG2 N5031 (L2/19-064), we propose to disunify nine Tangut ideographs and six Tangut components, as shown in Table 1 and Table 3. Proposed glyph changes for existing ideographs (including one pair of already disunified ideographs) are shown in Table 2.

In order to minimize disruption to existing Tangut data, for each disunified pair of ideographs the more common meaning remains attached to the existing code point, and the less common meaning is transferred to the new code point. This means that existing Unicode Tangut data using very frequently-occurring characters such as “heaven” (184F1) and “big” (18736) will not be affected by the disunification. In each case the source reference identifies the meaning of the ideograph.

As there are only eight free code points in the Tangut block, and nine new characters are proposed for encoding, it is necessary to create a new Tangut Extended (or Tangut Supplement) block. It is anticipated that in the future there will be occasional proposals to encode a small number of additional Tangut ideographs, but as minor glyph variations should be dealt with by means of Ideographic Variation Sequences, we think that an additional block of 128 code points should be sufficient for Tangut encoding needs. Therefore we suggest allocating a Tangut Extended block at 18D00..18D7F. As it is helpful to keep the new ideographs proposed in this document together, we propose to encode them in the Tangut Extended block at 18D00 through 18D08.

Taking into account the seven Tangut components already under consideration (see WG2 N4957; L2/18-194), there are six free code points in the Tangut Components block, so the six components proposed in this document can be added at 18AFA through 18AFF.

Table 1: Proposed Disunifications of Tangut Ideographs

Existing Ideographs			New Ideographs		
C/P	New Glyph*	Source Reference and Meaning	C/P	Glyph	Source Reference and Meaning
17134	𐞪	L2008-3488-3489 L2008-3488 (“pair”)	18D00	𐞪	L2008-3489 (“foolish, stupid”)
175F6	𐞫	L2008-1666-1667 L2008-1666 (“fox”)	18D01	𐞫	L2008-1667 (“tail, east”)
17F0D	𐞬	L2008-3435-3436 L2008-3436 (“close relative”)	18D02	𐞬	L2008-3435 (“god”)
17F8A	𐞭	L2008-2252-2253 L2008-2253 (“warehouse”)	18D03	𐞭	L2008-2252 (“kind of bird”)
17FA5	𐞮	L2008-3683-3684 L2008-3683 (“day after tomorrow”)	18D04	𐞮	L2008-3684 (“kind of bird”)
18139	𐞯	L2008-1317-1318 L2008-1317 (“to brush”)	18D05	𐞯	L2008-1318 (“to jump”)
18147	𐞰	L2008-1734-1735 L2008-1734 (negative prefix)	18D06	𐞰	L2008-1735 (“respectful”)
184F1	𐞱	L2008-1106-1107 L2008-1107 (“heaven”)	18D07	𐞱	L2008-1106 (“swallow”)
18736	𐞲	L2008-4456-4457 L2008-4457 (“big”)	18D08	𐞲	L2008-4456 (“wild goose”)

* The five highlighted glyphs are changed to use the new 𐞫 component instead of the 𐞫 component. See Table 2 for detailed explanation of proposed glyph changes.













Table 2: Proposed Glyph Changes for Tangut Ideographs

C/P	Old Glyph	New Glyph	Notes
17134	𗇑	𗇑	No change
175F6	𗇒	𗇒	Cosmetic change to 𗇒 component
17F0D	𗇓	𗇓	Change 𗇓 component to 𗇓 component
17F8A	𗇔	𗇔	Change 𗇓 component to 𗇓 component
17FA5	𗇕	𗇕	Change 𗇓 component to 𗇓 component
180D6	𗇖	𗇖	Change 𗇓 component to 𗇓 component Change 𗇓 component to 𗇓 component
18139	𗇗	𗇗	Cosmetic change to 𗇒 component
18147	𗇘	𗇘	Cosmetic change to 𗇒 component
182F5	𗇙	𗇙	No change
184F1	𗇚	𗇚	Change 𗇓 component to 𗇓 component
18736	𗇛	𗇛	Change 𗇓 component to 𗇓 component

Notes on glyph changes

1. For 17F0D, 17F8A, 17FA5, 184F1, and 18736, the existing glyph form (which represents the less common meaning) is transferred to the corresponding new character, and the existing character has a new glyph form with the 𗇓 component (which represents the more common meaning).
2. For 175F6, 18139, and 18147, the 𗇒 component has been slightly modified in order to accentuate the difference between it and the corresponding new ideograph with the 𗇒 component.
3. 180D6 (𗇖) “fountainhead” and 182F5 (𗇙) “vulture” are currently distinguished by their left side component (180D6 has Component 301 𗇓, whereas 182F5 has Component 412 𗇓). However, their left side components should both be Component 412, and the distinction between the characters should be the middle component, which should be 𗇓 for 182F5 and 𗇓 for 180D6.

Table 3: Proposed Disunifications of Tangut Components

Existing Components			New Components		
C/P	Glyph*	Character Name	C/P	Glyph	Character Name
18843		TANGUT COMPONENT-068	18AFA		TANGUT COMPONENT-763
18856		TANGUT COMPONENT-087	18AFB		TANGUT COMPONENT-764
1888C		TANGUT COMPONENT-141	18AFC		TANGUT COMPONENT-765
1890A		TANGUT COMPONENT-267	18AFD		TANGUT COMPONENT-766
18915		TANGUT COMPONENT-278	18AFE		TANGUT COMPONENT-767
1893B		TANGUT COMPONENT-316	18AFF		TANGUT COMPONENT-768

* For the existing components two glyph forms are shown. The first is the current code chart glyph, and the second is the proposed new code chart glyph, which has been slightly modified in order to accentuate the difference between it and the corresponding new component.

2. Unicode Properties

UCD properties for Tangut Components:

18AFA;TANGUT COMPONENT-763;Lo;0;L;;;;;N;;;;;
18AFB;TANGUT COMPONENT-764;Lo;0;L;;;;;N;;;;;
18AFC;TANGUT COMPONENT-765;Lo;0;L;;;;;N;;;;;
18AFD;TANGUT COMPONENT-766;Lo;0;L;;;;;N;;;;;
18AFE;TANGUT COMPONENT-767;Lo;0;L;;;;;N;;;;;
18AFF;TANGUT COMPONENT-768;Lo;0;L;;;;;N;;;;;

UCD properties for Tangut Extended:

18D00; TANGUT IDEOGRAPH-18D00;Lo;0;L;;;;;N;;;;;
18D01; TANGUT IDEOGRAPH-18D01;Lo;0;L;;;;;N;;;;;
18D02; TANGUT IDEOGRAPH-18D02;Lo;0;L;;;;;N;;;;;
18D03; TANGUT IDEOGRAPH-18D03;Lo;0;L;;;;;N;;;;;
18D04; TANGUT IDEOGRAPH-18D04;Lo;0;L;;;;;N;;;;;
18D05; TANGUT IDEOGRAPH-18D05;Lo;0;L;;;;;N;;;;;
18D06; TANGUT IDEOGRAPH-18D06;Lo;0;L;;;;;N;;;;;
18D07; TANGUT IDEOGRAPH-18D07;Lo;0;L;;;;;N;;;;;
18D08; TANGUT IDEOGRAPH-18D08;Lo;0;L;;;;;N;;;;;

Data for TangutSrc.txt (ISO/IEC 10646) and TangutSources.txt (UCD)

Modified entries:

U+17134 kTGT_MergedSrc L2008-3488
U+17134 kRSTUnicode 17.7
U+175F6 kTGT_MergedSrc L2008-1666
U+175F6 kRSTUnicode 87.9
U+17F0D kTGT_MergedSrc L2008-3436
U+17F0D kRSTUnicode 262.10
U+17F8A kTGT_MergedSrc L2008-2253
U+17F8A kRSTUnicode 267.9
U+17FA5 kTGT_MergedSrc L2008-3683
U+17FA5 kRSTUnicode 267.11
U+180D6 kTGT_MergedSrc L2008-5191
U+180D6 kRSTUnicode 412.14
U+18139 kTGT_MergedSrc L2008-1317
U+18139 kRSTUnicode 316.10
U+18147 kTGT_MergedSrc L2008-1734
U+18147 kRSTUnicode 316.11
U+184F1 kTGT_MergedSrc L2008-1107
U+184F1 kRSTUnicode 485.12

U+18736 kTGT_MergedSrc L2008-4457
U+18736 kRSTUnicode 674.14

New entries:

U+18D00 kTGT_MergedSrc L2008-3489
U+18D00 kRSTUnicode 17.7
U+18D01 kTGT_MergedSrc L2008-1667
U+18D01 kRSTUnicode 87.9
U+18D02 kTGT_MergedSrc L2008-3435
U+18D02 kRSTUnicode 262.10
U+18D03 kTGT_MergedSrc L2008-2252
U+18D03 kRSTUnicode 267.9
U+18D04 kTGT_MergedSrc L2008-3684
U+18D04 kRSTUnicode 267.11
U+18D05 kTGT_MergedSrc L2008-1318
U+18D05 kRSTUnicode 316.10
U+18D06 kTGT_MergedSrc L2008-1735
U+18D06 kRSTUnicode 316.11
U+18D07 kTGT_MergedSrc L2008-1106
U+18D07 kRSTUnicode 485.12
U+18D08 kTGT_MergedSrc L2008-4456
U+18D08 kRSTUnicode 674.14

Note on radicals and stroke counts

The proposed radical and stroke count values for the new disunified ideographs are the same as for the corresponding existing ideographs. This is because the kRSTUnicode key reflects the principles for assigning radicals and stroke counts that are defined in the original Tangut encoding proposal (WG2 N4522; L2/14-023). This means that sorting algorithms using kRSTUnicode will sort the new ideographs together with the corresponding existing ideographs, but this is consistent with the current situation for all ideographs ordered under Component 267 (𐫡) and Component 316 (𐫢), which in both cases comprise a mixture of characters which should be separated into two different radicals according to the new research of Jia and Jing.

New radical assignments and stroke counts for these characters should be part of additional radical/stroke data to be added as a new key to the Tangut data files at a future date. This new key would reflect the radical assignments and stroke counts provided in the final published version of Jia and Jing's *Xīxià zìfú jí shǔxìng biāozhù biǎo* (西夏字符及属性标注表) or other appropriate published document.

A new code chart font (or a revised version of the current code chart font) which uses glyph forms which correspond to the radicals and stroke counts assigned in the new radical/stroke key should be provided at the same time that the new radical/stroke data is added to the data files.

3. Proposal Summary Form

SO/IEC JTC 1/SC 2/WG 2 PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹. 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:	<i>Proposal to encode nine Tangut ideographs and six Tangut components</i>
2. Requester's name:	<i>Andrew West, Viacheslav Zaytsev, Jia Changye, Jing Yongshi, Sun Bojun</i>
3. Requester type (Member body/Liaison/Individual contribution):	<i>Individual contribution</i>
4. Submission date:	<i>2019-05-27</i>
5. Requester's reference (if applicable):	
6. Choose one of the following:	
This is a complete proposal:	<input type="checkbox"/> YES
(or) More information will be provided later:	<input type="checkbox"/>

B. Technical – General

1. Choose one of the following:	
a. This proposal is for a new script (set of characters):	<input type="checkbox"/> NO
Proposed name of script:	
b. The proposal is for addition of character(s) to an existing block:	<input type="checkbox"/> YES
Name of the existing block:	<i>Tangut and Tangut Components</i>
2. Number of characters in proposal:	<i>15</i>
3. Proposed category (select one from below - see section 2.2 of P&P document):	
A-Contemporary <input type="checkbox"/> B.1-Specialized (small collection) <input type="checkbox"/> B.2-Specialized (large collection) <input type="checkbox"/>	
C-Major extinct <input type="checkbox"/> D-Attested extinct <input checked="" type="checkbox"/> E-Minor extinct <input type="checkbox"/>	
F-Archaic Hieroglyphic or Ideographic <input type="checkbox"/> G-Obscure or questionable usage symbols <input type="checkbox"/>	
4. Is a repertoire including character names provided?	<input type="checkbox"/> YES
a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?	<input type="checkbox"/> YES
b. Are the character shapes attached in a legible form suitable for review?	<input type="checkbox"/> YES
5. Fonts related:	
a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?	<i>Jing Yongshi and Andrew West</i>
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):	<i>Jing Yongshi</i>
6. References:	
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?	<i>See WG2 N5031</i>
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?	<i>See WG2 N5031</i>
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)?	<input type="checkbox"/> 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: 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? If YES explain	<i>NO</i>
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.)? If YES, with whom? If YES, available relevant documents:	<i>Yes</i> <i>Tangut experts</i> <i>WG2 N4736</i>
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? Reference:	<i>NO</i>
4. The context of use for the proposed characters (type of use; common or rare) Reference:	<i>Rare</i>
5. Are the proposed characters in current use by the user community? If YES, where? Reference:	<i>YES</i>
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? If YES, is a rationale provided? If YES, reference:	<i>NO</i>
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	<i>YES</i>
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? If YES, is a rationale for its inclusion provided? If YES, reference:	<i>NO</i>
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? If YES, is a rationale for its inclusion provided? If YES, reference:	<i>NO</i>
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? If YES, is a rationale for its inclusion provided? If YES, reference:	<i>YES</i> <i>YES</i> <i>WG2 N5031</i>
11. Does the proposal include use of combining characters and/or use of composite sequences? 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:	<i>NO</i>
12. Does the proposal contain characters with any special properties such as control function or similar semantics? If YES, describe in detail (include attachment if necessary)	<i>NO</i>
13. Does the proposal contain any Ideographic compatibility characters? If YES, are the equivalent corresponding unified ideographic characters identified? If YES, reference:	<i>NO</i>