JTC1/SC2/WG2 N4033
2011-05-22

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

Doc Type:	**Working Group Document**
Title:	**Report on Tangut Encoding**
Source:	**UK**
Status:	**National Body Contribution**
Action:	**For consideration by JTC1/SC2/WG2 and UTC**
Date:	**2011-05-22**

---

1. Summary

N3797 ("Final proposal for encoding the Tangut script in the SMP of the UCS") submitted by China, Ireland, and UK was considered at WG2 M56, and a number of issues were raised at the ad hoc meeting (see N3833). Resolution M56.17 requested that a revised proposal taking into account the ad hoc recommendations be submitted. The UK has carried out a thorough review of the proposed repertoire, collating the 6,055 proposed characters in N3797 against other important sources (Li Fanwen 1986, Han Xiaomang 2004, Kychanov and Arakawa 2006), and we request that a revised repertoire of 6,080 characters, as detailed in this document, be added to a new amendment at the earliest opportunity. We believe that this revised repertoire is as complete as can practically be expected, and we do not anticipate the need for a supplementary block of Tangut characters.

2. Encoding Principles

It is difficult to define the repertoire of Tangut characters to be encoded, as there is no single contemporary source that lists all possible characters. Tangut ideographs are often written with different glyph forms in different sources, and modern Tangut dictionaries include entries for many of the more important variant glyph forms. Because of the complexity of the characters, and the often poor quality of the primary source texts, modern scholars have frequently misinterpretted or miswritten the original Tangut characters, with the result that modern dictionaries include some ghost characters, and the glyph form of the same character often varies significantly from one dictionary to another (a stroke added or omitted, or one component substituted for another component). It is neither practical nor desirable to encode all the attested variant glyph forms, but scholars do need to be able to represent all Tangut characters that have been identified as characters in modern dictionaries, and they often need to discuss character variants (see for example http://www.amritas.com/110521.htm#05152333 where Tangutologist Marc Miyake notes how "small typographical errors can lead a researcher astray"). In particular, the Mojikyo set of 6,000 Tangut characters corresponding to the repertoire in Li Fanwen 1997 has been very widely used by Tangutologists, and it is important to be able to easily convert Tangut text written using the Mojikyo set into Unicode Tangut.

We therefore propose to unify characters with different glyph forms in different source dictionaries where the glyph variants are not used contrastively in the same source (i.e. where Source A represents character C1 with glyph variant V1, and Source B represent the same character with glyph variant V2, but no source gives both V1 and V2 as separate characters, then V1 and V2 are unified). It is appropriate to represent such glyph variants using variation sequences if it is required to indicate the differences at the encoding level. If this encoding proposal is accepted we will put forward a proposal for a set of standardized variation sequences for significant character variants that have been unified in this proposal.

On the other hand, where a source dictionary gives two or more glyph variants of the same character as separate entries then we propose that they be encoded separately, as it is not appropriate for users to have to use variation sequences to represent individual head characters in a single dictionary, especially in the case of recent authoritative dictionaries such as Kychanov & Arakawa 2006 and Li Fanwen 2008. However, we do not expect that this principle should apply to any future dictionaries, and that any variants unified in the current proposal would not automatically be candidates for encoding if they were to be given separate entries in a future dictionary (of course they may become candidates for encoding if a semantic difference between the variants could be demonstrated). Likewise, we do not apply this principle to older and less authoritative sources that have not been used as primary sources for the character repertoire (in any case, most of the character variants given in earlier works are included in LFW2008).

The proposed character repertoire is derived from the following sources:

- **HXM2004**. Hán Xiǎománg (韓小忙). 西夏文正字研究 (= *Xīxiàwén Zhèngzì Yánjiū*). 2004.
- **KYC2006**. Kychanov, E. I. (Е. И. Кычанов) and Arakawa Shintarō 荒川 慎太郎. *Словарь тангутского (Си Ся) языка* (= *Slovar' tangutskogo (Si Sja) jazyka*) [Tangut-Russian-English-Chinese Dictionary]. St. Petersburg and Kyoto, 2006.
- **LFW1986**. Lǐ Fànwén (李範文). 同音研究 (= *Tóngyīn Yánjiū*). Yinchuan, 1986.
- **LFW1997**. Lǐ Fànwén (李範文). 夏漢字典 (= *Xià-Hàn Zìdiàn*). Beijing, 1997.
- **LFW2008**. Lǐ Fànwén (李範文). 夏漢字典 (= *Xià-Hàn Zìdiàn*). Beijing, 2008.

In the case of HXM2004, which as an unpublished dissertation has less authority and is less widely used than the other sources, we only propose that characters given in the final column ("character types") are encoded, and that glyph variants given in the first column ("character variants") are unified if they are not given separate entries in one of the other sources (there are 135 such unified variants from HXM2004 column 1).

3. Proposed Repertoire

We propose to encode a repertoire of 6,080 characters, comprising 6,079 Tangut ideographs and one iteration mark, as listed in Appendix A of this document. This is a superset of the repertoires proposed in N3297 and N3797. The repertoire comprises the 6,055 characters from LFW2008 proposed in N3797, plus 25 additional characters from other important sources:

Additions for HXM2004 (10 characters)

- #1015 (N3297 U+17419) = U+17442
- #1805 (N3297 U+1773B) = U+177A0
- #5647 (N3297 U+18686) = U+186D0

- #5796 (N3297 U+182A8) = U+18362
- #5829 (N3297 U+17373) = U+1738D
- #5839 (N3297 U+17D03) = U+1785A
- #5841 (N3297 U+17D0A) = U+17D58
- #5844 (N3297 U+17D47) = U+17D90
- #5845 (N3297 U+17D4E) = U+17D97
- #5854 (N3297 U+180D4) = U+1815A

Additions for LFW1986 (3 characters)

- #1742 (also given as #6000 in the index of LFW1997) = U+17B3A
- #4840 = U+1841E
- #5541 = U+184CC

Additions for LFW1997 (5 characters)

- #5995 = U+18671
- #5996 = U+17857
- #5997 = U+18126
- #5998 = U+178F3
- #5999 = U+17998

Additions for LFW2008 (6 characters)

- #2941B = U+17056
- #1267B = U+17204
- #3007B = U+172F9
- #3286B = U+179F2
- #5150B = U+17B5B
- #2799B = U+17F2F

Additions for KYC2006 (1 character)

- #4068 = U+179F0

The following changes and corrections to the order of characters given in N3797 have also been implemented in the new repertoire.

(i) Corrections to N3797 IDS sequences and/or sort key:

- 17342 (L0325) stroke count corrected from 23 to 13, and reordered between 17333 and 17334
- 175E0 (L1311) stroke count corrected from 11 to 10, and reordered between 175DD and 175DE

- 17AB0 (L3147) stroke count corrected from 24 to 13, and reordered between 179E4 and 179E5
- 17E17 (L1986) IDS sequence corrected, but no reordering required
- 18163 (L1046) IDS sequence corrected, but no reordering required
- 18670 (L2690) stroke count corrected from 14 to 13, and reordered between 1866A and 1866B
- 186AC (L5552) sort key corrected, but no reordering required
- 186C6 (L0420) stroke count corrected from 14 to 15, and reordered between 186C7 and 186C8

(ii) Reordering of characters with the same sort key that differ by final stroke realization to accord with the sorting principles given in N3797 (character with bent final stroke should go after a similar character with straight final stroke):

- 17051 (L1975) reordered before 17050 (L1960)
- 1708A (L3496) reordered before 17089 (L3462)
- 170B1 (L2987) reordered before 170B0 (L2922)
- 171C8 (L3841) reordered before 171C7 (L3837)
- 173FE (L3820) reordered before 173FD (L3814)
- 1743F (L4943) reordered before 1743E (L4863)
- 1745F (L4951) reordered before 1745E (L4862)
- 174BE (L4944) reordered before 174BD (L4864)
- 175A9 (L4069) reordered before 175A8 (L4066)
- 1761E (L0202) reordered before 1761D (L0162)
- 176AF (L4318) reordered before 176AE (L4212)
- 176F2 (L4415) reordered before 176F1 (L4408)
- 1780D (L4534) reordered before 1780C (L4514)
- 178A2 (L3191) reordered before 178A1 (L3114)
- 178B0 (L2144) reordered after 178B2 (L2290)
- 178D8 (L2748) reordered before 178D7 (L2089)
- 178DD (L3200) reordered before 178DC (L2189)
- 178F4 (L3691) reordered before 178F3 (L3590)
- 179EA (L3381) reordered before 179E9 (L3154)
- 179F6 (L3260) reordered before 179F5 (L3175)
- 17A69 (L3326) reordered before 17A68 (L3145)
- 17BBF (L5240) reordered before 17BBE (L5185)
- 17C3F (L1414) reordered before 17C3E (L1208)
- 17D9D (L1665) reordered before 17D9C (L1605)
- 17E7F (L2687) reordered before 17E7E (L2622)
- 17E8F (L2508) reordered before 17E8E (L2502)
- 17F8B (L6032) reordered before 17F8A (L2451)
- 18019 (L6061) reordered before 18018 (L5310)
- 1820B (L3447) reordered before 1820A (L3444)
- 18237 (L2059) reordered before 18236 (L2010)
- 18305 (L5152) reordered before 18304 (L5096)
- 18311 (L6060) reordered before 1830F (L5263)
- 1833C (L6062) reordered before 1833B (L5311)
- 18352 (L5009) reordered before 18351 (L4912)
- 1838B (L5246) reordered before 1838A (L5199)
- 183CD (L5404) reordered before 183CC (L5373)
- 183EC (L5764) reordered before 183EB (L5729)
- 18519 (L5925) reordered before 18518 (L5920)
- 185EC (L0385) reordered before 185EA (L0261)

- 1867A (L6030) reordered before 18679 (L2402)

(iii) Unification of Radicals R349 and R350:

- 18588 and 18589 reordered between 18563 and 18564
- 1858A between reordered 18564 and 18565
- 1858B between reordered 18566 and 18567
- 1858C between reordered 18567 and 18568
- 1858D between reordered 18578 and 18579

4. Allocation

It is proposed that the characters be encoded in a block entitled "Tangut Ideographs" at 17000 through 187FF, with the Tangut Iteration Mark at 17000, and the 6,079 Tangut ideographs at 17001..187BF. This would leave 64 reserved characters at the end of the block (187C0..187FF), which we believe should be sufficient for any additional characters that may need to be encoded in the future (there is a large corpus of Tangut texts, and so it is quite possible that new characters may be identified in the future).

5. Character Names

As agreed at M56, the Tangut ideographs should be named algorithmically "TANGUT IDEOGRAPH-17001" through "TANGUT IDEOGRAPH-187BF", but the Tangut iteration mark (U+17000) should be named "TANGUT ITERATION MARK".

6. Outstanding Issues

Whilst reviewing the repertoire, we have noticed some characters in LFW2008 (and consequently in the font to be used in the code charts, which was used to typeset LFW2008) that may be drawn incorrectly. For example, LFW2008 #0949 (= U+17C78) is drawn with 16 strokes, whereas all other sources draw this character with 17 strokes; checking the facsimile reprint of the *Sea of Characters* (*Wénhǎi* 文海) it does indeed appear that it should be drawn with 17 strokes, and that the glyph form given in LFW2008 is a mistake. We would therefore like to request that the Chinese NB review the glyph forms of all characters in LFW2008 (in particular those characters that differ from other sources) against the original Tangut sources, to ensure that the representative glyphs are correct and the characters are all ordered correctly. However, we do not think this issue should be an impediment to putting the Tangut repertoire proposed in this document in a new amendment at the earliest opportunity, as any problems with the glyphs or character ordering can be resolved in ballot comments.

7. Appendices

Two appendices are attached to this document. The information given in these appendices will be made available in Excel or plain text format to national bodies or interested experts upon request.

Appendix A lists the characters to be encoded, together with IDS sequences, radical, stroke count and sort key for each Tangut ideograph. Where a character unifies two or more glyph variants from different sources, multiple IDS sequences are provided on separate lines: the first IDS sequence coresponds to the code chart glyph, and subsequent IDS sequences correspond to unified glyph variants. The actual glyph variants for the character can be seen in Appendix B.

Appendix B is a multi-column chart showing the glyph forms and reference numbers for various sources. The following information is provided:

- **Code Point** : proposed code point
- **Glyph** : proposed representative glyph derived from the font created by Jing Yongshi and provided by the Chinese NB
- **N3297** : glyph and code point given in the SEI proposal authored by Richard Cook (WG2 N3297)
- **HXM2004** : glyph and reference numbers from HXM2004, where the first number is the "character type" (last column), and the second number is the "character variant" (first column)
- **Mojikyo** : glyph and reference number from the Mojikyo fonts "Mojikyo M202" (version 3.0) and "Mojikyo M203" (version 3.0)
- **LFW1997** : glyph and reference number from LFW1997
- **LFW2008** : glyph and reference number from LFW2008 (LFW2008 is typeset using the Jing Yongshi font that is to be used for the code charts)
- **KYC2006** : glyph and reference number from KYC2006
- **LFW1986** : glyph and reference number from the radical/stroke index of LFW1986