ISO/IEC JTC1/SC2/WG2
Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Title: Proposal to encode Tangut characters in UCS Plane 1

Doc Type: Working Group Document
Source: UC Berkeley Script Encoding Initiative (Universal Scripts Project)
Author: Richard COOK
Status: Liaison Contribution
Action: For consideration by JTC1/SC2/WG2 and UTC
Date: 2007-05-09

This proposal presents a new block of 5,910 Tangut (a.k.a. 西夏, Xī Xià, Тангут, Си Ся) characters for encoding in Plane 1 of the UCS, in the range (U+17000..U+18715), coordinating contributions from scholars in China, Japan, Russia, Taiwan, and the United States.

Tangut is an extinct language of central China (都興慶府, 今寧夏銀川). Tangut characters were in use for less than 500 years (1036-1502), including some 300 years of classical use beyond the Mongol destruction of Tangut civilization (1227). Tangut writing was invented by imperial decree, with Chinese writing as its conceptual model. Tangut characters are confined to a uniform em-square, and comprised of a finite set of stroke types combined in recurrent patterns. The resemblance to Chinese and CJK writing ends there: Tangut is a unique writing system, with no overlap with Unified CJK Ideographs. There is a considerable body of native Tangut literature that has been uncovered since the early 1900s, and a large body of paleographic and linguistic work has also been published, making a standard encoding of great consequence for future research. The first attempt to define a standard electronic encoding of Tangut was Grinstead's "Tangut Telecode" (1971), but neither this nor more ambitious subsequent computerized systems have gained much currency among scholars, though a *de facto* standard point of reference in Tangut studies has slowly emerged.

The proposed repertory derives from three main sources, two by 李范文 *Lǐ Fànwén* (**TY**:1986, **XiaHan**:1997) and one by 韓小忙 *Hán Xiǎománg* (**HXM**:2004). *Hán Xiǎománg* was the executive editor of his teacher *Lǐ Fànwén's* mammoth 1997 《夏漢字典》*Xià-Hàn Zìdiǎn* [Tangut / Chinese Dictionary]. Seven years later Han's doctoral dissertation 《西夏文正字研究》 [On Tangut Orthography] presents a comprehensive and systematic analysis of the distinctive elements of Tangut writing, mapped to nine native Tangut dictionaries.

Complete mappings and glyphs for these three sources (TY, XiaHan, HXM) are provided in the *Multi-Column Code Chart* (**L2/07-144**) accompanying this proposal; these mappings are a subset of the complete *UniHan*-style Tangut mapping database, described in the *Proposed Draft Unicode Technical Report #43: A User's Guide to the UniTangut Database* <http://unicode.org/~rscook/Xixia/UCS_proposal/tr43.html>.

The *Multi-Column Code Chart* (**L2/07-144**) has four fields per record, one field per source, and 6217 records total (including variants), for 5910 proposed new characters. It is printed with 24 pt. Tangut glyphs, in 4 columns and 17 records per page, over 92 pages.

The four columns of the *Multi-Column Code Chart* have header labels "**W**" (TY:1986), "**X**" (XiaHan:1997), "**Y**" (HXM:2004), "**Z**" (proposed UCS representative glyph). Under each glyph in the first 3 columns (W,X,Y) is its source mapping, and under each cell in column Z is the proposed UCS code point. Each record in this chart has at least two glyphs, and variant classes occupy adjacent rows with the same proposed code point: the first row for each variant class provides the proposed representative glyph in field Z (as determined

by HXM).

The *Single-Column Code Chart* (**L2/07-145**) shows only the column "Z" proposed representative glyph.

The consolidated mappings and glyphs underwent rigorous proofing and correction over the course of the year 2006 and into 2007, resulting in the set of five related Tangut *TrueType* fonts presented in this proposal. The base font and mappings for column "W" came from Taiwan (Academia Sincia); the fonts for columns "X", and "Y" were created specifically for this proposal, based on scans of the two main print-sources (Xia-Han, HXM). The font for column "Z" and its subset font for the *Single-Column Code Chart* were also built specifically for this proposal, based on a bitmap set from Japan (created in work by 荒川慎太郎 *Arakawa Shintaro* et al., based also on the XiaHan source), corrected, expanded, and brought into line with the HXM source. Three of the fonts in the *Multi-Column Code Chart* (X,Y,Z) contain glyphs for the full proposal repertory (5910 characters); the Sinica font (W) only contains *Tongyin* (TY) glyphs.

The ordering in the *Single* and *Multi- Column Charts* derives from the system presented in HXM (2004), and virtual positions are assigned for characters not in that source. That scheme is especially attractive for its logic and high degree of usability. For *every* character, the left-side, top- or bottom-spanning *component* is the "radical". The stroke types and counts of these components order these component classes, and within a given class the ordering is also by stroke count and by the type of the first residual stroke. This ordering eliminates the immediate need for the encoding of a block of Tangut radicals. The competing systems of Tangut radicals (and there are several, though apparently no known native systems) are idiosyncratic and partial. The task of enumerating the complete set of Tangut radicals is rather open-ended, since this is in effect a subset of the similarly open-ended set of Tangut components (which is especially open-ended when character variants, and variant component analyses are considered). On-going CDL analysis of the encoded Tangut repertory will provide the basis for the future encoding of a well-defined set of Tangut components. CDL itself is the most effective way to address the radical/component problem in Tangut, and the larger problems relating to the indexing of this large character set.

Tangut characters are processed for most purposes like Chinese characters. They occupy unit squares, line-breaks can occur between any characters (subject to ordinary Han script rules for placement of punctuation). The default sort order is determined by binary (code-chart) order. Tangut, Chinese, Russian, English, IPA, etc. are commonly intermixed in the same lines of left-to-right running text. Tangut may also be used in vertical text. For examples, see the images at <http://unicode.org/~rscook/Xixia/>.

The *Single-Column Code Chart* (**L2/07-145**) includes at the end a full set of names. All character names have the form "`TANGUT CHARACTER-17000`". Due to the large number of Tangut characters, it is suggested that *UniHan*-style short-hand notation be used in "NamesList.txt".

In addition to lexical source mappings, the mapping data documented in *Proposed Draft Unicode Technical Report #43: A User's Guide to the UniTangut Database* will provide several other kinds of property information. Some of these are described below.

Sources and Properties

The following is a list of fields and abbreviations used in the online mapping database:

<http://linguistics.berkeley.edu/~rscook/cgi/ztangut.html>

B5 : Academia Sinica's Big5-based encoding of this character (see PUA).

HXM : 韓小忙 Hán Xiǎománg (2004): 《西夏文正字研究》 [*On Tangut Orthography*; Ph.D. dissertation K246.3 H211.7, directed by 李范文 Lǐ Fànwén, see TY]. HXM undertakes a comprehensive and systematic collation of Tangut characters, based on nine Tangut dictionaries (《同音》, 《文海寶韻》, 《同音文海寶韻合編》, 《番漢合時掌中珠》, 《三才雜字》, 《纂要》, 《同義》, 《五音切韻》, 《新集碎金置掌文》), and catalogues a total of 6,066 forms, including 169 variants, 36 errors, and 5,861 unique 'standard-style characters' ("正字" *zhèngzì* 'orthography'). In addition to the primary source mappings, this work contains mappings to Lǐ (1997) and Sofronov (1968).

Kychanov : Е. И. Кычанов *Словарь Тангутского (Си Ся) Языка* [E.I. Kychanov; *Tangut Dictionary: Tangut-Russian-English-Chinese Dictionary*], Kyoto Univ., 2006; this dictionary uses the Arakawa Mojikyo (文字鏡) fonts; pronunciations after Sofronov, gloss material after 李范文; 5803 indexed entries, including many variants.

Nevsky : Н. А. Невский (Nevsky, N. A.) [1892-1938] (1960) *Тангутская Филология*. Издатепьство Восточны литературы, Москва [*Tangut Philology*. 2 vols (Russian) Moskow]. (This field may have a maximum of four space-delimited values.)

Nishida : 《西夏語的研究》 (西田龍雄 Nishida Tatsuo) 第二冊, 西夏文字小字典 Appendix I, p. 303-507 上面的編號.

PUA : Academia Sinica's Unicode *Private Use Area* encoding of B5, see UNI.

SN : Serial Number, a numbering of all 5,809 elements of the TY character set.

Sofronov : Софронов, М. В. (M. V. Sofronov) 索夫羅諾夫著的《西夏語文法》 (*Грамматика Тангуцково Языка* [*Grammatika Tangutskovo Yazyka* 'Tangut Garammar'], 1968).

TY : 《同音研究》*Tóngyīn Yánjiū* ('Homophones' Research, 李范文 Lǐ Fànwén. 寧夏人民出版社, 1986). The Sinica TY database contains a total of 5,809 records, including variants.

TYBH : 《同音研究》筆畫, (TY Stroke-count Range).

TYBS : 《同音研究》部首 (TY Radical).

TYP : 《同音研究》品, (TY Class).

TYYY : 《同音研究》音韻, (TY Rhyme).

TYYZ : 《同音研究》頁字, (TY character mapping [page + character ID]).

UNI : non-*PUA* Unicode code point (in the proposed range [U+17000 .. U+18715]), with block ordering as in HXM.

WHYJ : 《文海研究》 *Wén Hǎi Yánjiū* (史金波, 白濱, 黃振華, 1983).

WenHai : 《文海》 *Wén Hǎi* (K. V. Keping et al., 1969).

XiaHan : 《夏漢字典》*Xià-Hàn Zìdiǎn* [Tangut / Chinese Dictionary] (李范文 Lǐ Fànwén 1997; ISBN: 7-5004-2113-3). This dictionary has 6,000 numbered entries, including a number of variants unified in the proposed repertory. In the mapping data, indices > 6000 are virtual.

YTYL : 《義同》一類 *Yì Tóng* yīlèi (李范文, 韓小忙, 2000; cf. 韓小忙 2004:354).

Several other fields are in the process of being added to this database, including phonological and semantic information. For additional information on Tangut and on the mapping database, see the document *Research Notes*:

<http://unicode.org/~rscook/Xixia/>

Acknowledgements

The astonishing labor of 李范文 Lǐ Fànwén (1997, 1986) and 韓小忙 Hán Xiǎománg (2004) made this proposal possible. Their sustained work served as the primary basis for determining the repertory and ordering.

For assistance in preparation of the mapping and font data for the present proposal, I am indebted to the following researchers at 中央研究院語言學研究所 *Linguistics Institute, Academia Sinica*, Taiwan: 龔煌城 GONG Hwang-cherng, 高雅琪 GAU Yeachyi, 莊德明 CHUANG Derming, 鄭錦全 C.C. CHENG, 林英津 LIN Ying-chin, and 余文生 (Jonathan EVANS). Special thanks to 林英津 LIN Ying-chin for first teaching me about, and then lending me her copy of 《西夏文正字研究》.

Special thanks to 荒川慎太郎 ARAKAWA Shintaro and the *Mojikyo Institute*, for granting permission to use their bitmap data.

Inadequate thanks to Jim MATISOFF for lending me his copy of《同音研究》and for giving me the perfect research environment in which to enjoy it. Thanks to 池田巧 IKEDA Takumi for copies of《夏漢字典》and *Словарь Тангутского (Си Ся) Языка*, to Martin HEIJDRA (Princeton) for suggestions during the drafting of the online version of the *Research Notes*, to 魏安 (Andrew WEST) for inputting a large quantity of R/S index data (LFW 1997:1088-1166), and to 阿南康宏 ANAN Yasuhiro, and 加藤昌彦 KATO Atsuhiko for suggestions and information relating to the *Mojikyo* (文字鏡) Tangut glyphs.

Thanks to Ken LUNDE (Adobe.com) for supra-BMP *OpenType*, to Tom BISHOP (Wenlin.com) for his preëminent 文林 Unicode text and CDL font editor (may we all live so long to see CDL applied to Tangut!), and to George WILLIAMS for his most excellent *FontForge*, used to create the *TrueType* fonts for this proposal. Thanks to Peter SELINGER for auto-tracing (*Potrace*; Dalhousie Univ., Math) and Thorsten LEMKE (Lemke Software GmbH) for image processing. And thanks to Prof. Unicode (now Emeritus, a.k.a. Asmus FREYTAG) for offering a few moments to hash out the subtleties of *UniBook* in between marathon ed-com filibusters. And finally, thanks to Debbie ANDERSON (SEI) and Rick MCGOWAN, for believing in the viability of the plan to bring Tangut back from one thousand years of extinction.

This proposal was created (for the most part) on an Apple Macintosh PowerBook G4, in Mac OS X (10.4.9), using a lot of commercial and free software, including (but not limited to) *Perl* (v5.8.6 for darwin, and v5.8.4 for i386-linux), *X11* (1.1.3 - XFree86 4.4.0), *FontForge* (v20:02), *GraphicConverter* (5.9.4), *Potrace* (1.7.darwin6.0), *Wenlin* (3.3.10), Adobe *InDesign CS2* (4.0.4), *Acrobat* (7.0, 8.0), *Rsync* (2.6.3), *MySQL* (server versions 3.23.56 and 4.1.11), *FileMaker Pro* (8.0v1). The code charts were output using the unimitable *UniBook* (5.0) on Windows XP.

This proposal was made possible in part by grants from the U.S. *National Endowment for the Humanities* (NEH; PA-50709) and the *National Science Foundation* (BCS-0345929) to U.C. Berkeley's *Sino-Tibetan Etymological Dictionary and Thesaurus Project* (<http://stedt.berkeley.edu>), and by an NEH grant to the *Universal Scripts Project* (part of the *Script Encoding Initiative* <http://www.linguistics.berkeley.edu/sei/>).

OBLIGATORY DISCLAIMER: Opinions, findings, conclusions, and/or recommendations expressed in this proposal are those of the proposers, and do not necessarily reflect the views of any funding agencies.

A. Administrative

1. TITLE:

Proposal to encode Tangut characters in UCS Plane 1

2. REQUESTERS NAME(S):

SEI / Richard COOK.

3. REQUESTER TYPE:

Liaison Contribution.

4. SUBMISSION DATE:

2007-05-09.

5. REQUESTER'S REFERENCE

N.A.

6A. COMPLETION

This is a final proposal.

6B. MORE INFORMATION TO BE PROVIDED?

No.

B. Technical - General

1A. NEW SCRIPT? NAME?

Yes. Tangut.

1B. ADDITION OF CHARACTERS TO EXISTING BLOCK? NAME?

No.

2. NUMBER OF CHARACTERS

5,910.

3. PROPOSED CATEGORY

Category F.

4. PROPOSED LEVEL OF IMPLEMENTATION AND RATIONALE

Level 1.

5A. CHARACTER NAMES INCLUDED IN PROPOSAL?

Yes.

5B. CHARACTER NAMES IN ACCORDANCE WITH GUIDELINES?

Yes.

5C. CHARACTER SHAPES REVIEWABLE?

Yes (see below).

6A. WHO WILL PROVIDE COMPUTERIZED FONT?

Richard Cook.

6B. FONT CURRENTLY AVAILABLE?

Yes.

6C. FONT FORMAT?

TrueType.

7A. ARE REFERENCES (TO OTHER CHARACTER SETS, DICTIONARIES, DESCRIPTIVE TEXTS, ETC.) PROVIDED?

Yes, mapping table (Proposed Draft UTR), scans, and *Bibliography*.

7B. ARE PUBLISHED EXAMPLES (SUCH AS SAMPLES FROM NEWSPAPERS, MAGAZINES, OR OTHER SOURCES) OF USE OF PROPOSED CHARACTERS ATTACHED?

Yes, see mapping table and *Bibliography*.

8. DOES THE PROPOSAL ADDRESS OTHER ASPECTS OF CHARACTER DATA PROCESSING?

Yes.

C. Technical -- Justification

1. CONTACT WITH THE USER COMMUNITY?

Yes.

2. INFORMATION ON THE USER COMMUNITY?

Scholarly community.

3A. THE CONTEXT OF USE FOR THE PROPOSED CHARACTERS?

Used to write the Tangut (西夏 *Xi Xia*) language.

3B. REFERENCE

See *Bibliography*.

4A. PROPOSED CHARACTERS IN CURRENT USE?

Yes

4B. WHERE?

By scholars worldwide.

5A. CHARACTERS SHOULD BE ENCODED ENTIRELY IN BMP?

No. Positions [U+17000] .. [U+18715] in Plane 1 are proposed for the characters.

5B. RATIONALE

See Roadmap.

6. SHOULD CHARACTERS BE KEPT IN A CONTINUOUS RANGE?

Yes.

7A. CAN THE CHARACTERS BE CONSIDERED A PRESENTATION FORM OF AN EXISTING CHARACTER OR CHARACTER SEQUENCE?

No.

7B. WHERE?

N.A.

7C. REFERENCE

N.A.

8A. CAN ANY OF THE CHARACTERS BE CONSIDERED TO BE SIMILAR (IN APPEARANCE OR FUNCTION) TO AN EXISTING CHARACTER?

No.

8B. WHERE?

N.A.

8C. REFERENCE

N.A.

9A. COMBINING CHARACTERS OR USE OF COMPOSITE SEQUENCES INCLUDED?

No.

9B. LIST OF COMPOSITE SEQUENCES AND THEIR CORRESPONDING GLYPH IMAGES PROVIDED?

No.

10. CHARACTERS WITH ANY SPECIAL PROPERTIES SUCH AS CONTROL FUNCTION, ETC. INCLUDED?

No.