ISO/IEC JTC 1/SC 2/WG 2 N4094
Date: 2011-06-01

DOC TYPE: Working Group Document
TITLE: Comments on Tangut report N4033
SOURCE: Richard Cook and Deborah Anderson, Script Encoding Initiative, UC Berkeley
STATUS: Liaison member contribution
ACTION: For consideration by WG2
DISTRIBUTION: ISO/IEC JTC1/SC2/WG 2

The following are comments and suggestions for progressing Tangut, based on N4033:

(1) N4033 focuses on multi-column chart production, which is a very positive direction, but the chart does not identify the source of the fonts for each column. This information should be added into the proposal. (Richard Cook will provide fonts as needed for review of font changes in multi-column charts.)

(2) N4033 adds partial mappings to code charts in N3297, but no mention is made to either N3822 or N3521. Using the LFW1997 serial numbers, including the virtual mappings, would be helpful. See the datafile to N3521: http://linguistics.berkeley.edu/~rscook/UTC/Tangut/20081010/08349-n3521-data.txt.

Can explicit mappings (and their sources) be added to the charts? (Note that N3629, the 2009 Tangut ad hoc report, had requested: “Provide mapping data as requested by reviewers. Include additional sources and mappings to N3521.”)

(3) One of the biggest problems with the repertory is "LFW2008", as noted by the issues raised in section 6 of N4033, and we agree China should review the LFW2008 glyphs. More generally, review and revision of the proposal would benefit from active involvement of China and Russia (since their experts are responsible for a great deal of primary and secondary source data, and font data).

As part of their participation, China and Russia could provide images of the original manuscripts (as they both seem to be working on this) which could be tied into the database. This would give more confidence in the coverage of the repertory, suitability of representative glyphs, and help resolve the variant issue.

(4) Inclusion of the radicals in the repertory still seems questionable. The set may be idiosyncratic, and may not be comprehensive. Component lists may be subjective and open-ended, reflecting glyph-level rather than character-level features.

The following information, as requested in N3629, the 2009 Tangut ad hoc report, would be useful:

> “Provide information regarding the system of radicals used in N3577, including the basis on which it was devised, how it compares to other systems of radicals used by experts, and what implications for experts might be if it is used as the basis of ordering the repertoire in the UCS.”

Please justify the inclusion of radicals, components, and strokes (if any) in the main repertory. Are they authentic Tangut characters, or simply modern inventions? If they are characters suitable for UCS encoding, why should they be integrated into the main block, and not encoded in a separate block? The radical set raises a number of issues, and has the potential to greatly complicate determination of and agreement upon a repertory.

(5) Our original documents sought to address variants (via VS mechanism). The current proposal discusses them in section 2, saying “it is appropriate to represent such glyph variants using variation sequences if required to indicate the differences at the encoding level” and proposes putting forward a proposal if N4033 is accepted. In order to account for the full set of Tangut source mappings, explicit VS assignments should be included as part of the Tangut proposal.

(6) Our original proposal sought to compile/publish/maintain a public database of metadata (Proposed Draft Unicode Technical Report #43, accessible at: http://unicode.org/~rscook/Xixia/UCS_proposal/tr43.html). The current proposal contributes valuable metadata, but does nothing in regard to proposed TR #43. Suitable metadata for the full repertory should be made available (in plain text tab-delimited and/or XML form), for addition to/inclusion in the public repository.

(7) Given the valuable component and stroke data which UK provides, it would make sense to create a CDL database for multi-column Tangut, for proofing and to do the comprehensive treatment of radicals, strokes, components, and variants that seems to be required. This would probably take at least one year of concerted effort. We plan to work on this, and make the results available.

(8) The new character names are now algorithmic, but have "Ideograph" in the name. “Ideograph” in the UCS is used as a synonym for "CJK" and "Han", and should be avoided. The names should simply use "character" instead.

(9) We regard HXM as an important source of mapping and glyph data, providing key links to the original manuscripts. It should have highest priority in determination of the repertory.