L2/04-373

Document Type: Working Group Document
Title: Unencoded CJK Ideographs: Proposal to add kXHC data to Unihan.txt
Source: Richard Cook <rscook@unicode.org>
Status: Individual Contribution
Action: For consideration by JTC1/SC2/WG2, IRG, and UTC
Date: 2004-10-28

The important modern PRC lexical source 现代汉语词典 Xiandai Hanyu Cidian (Beijing: Shangwu Yinshu Guan, 1978 [1983, 1993], ISBN: 7-100-00044-0/H.18) contains a number of unencoded CJK ideographic characters.

A mapping table has been created to provide common pinyin readings for the ~ 14,000 character entries contained in this dictionary, including many common variant readings, and to provide an inventory of unencoded ideographs. This mapping table is in the final stages of proofing.

The following partial list gives the traditional characters, followed by a "?" indicating that the simplified form (or particular variant) appearing in this text is not currently encoded in the UCS.

䘞褭?   [U+461e][U+892d]?
雜杂?   [U+96dc][U+6742]?
彍? [U+5f4d]?
瓅? [U+74c5]?
顗? [U+9857]?
緺? [U+7dfa]?
薆? [U+8586]?
鮟? [U+9b9f]?
䶕? [U+4d95]?
韛? [U+97db]?
邦? [U+90a6]?
梖? [U+6896]?
屄? [U+5c44]?
閟? [U+959f]?
駜? [U+99dc]?
饆? [U+9946]?
鱍? [U+9c4d]?
賨? [U+8ce8]?

This partial list only contains items catalogued so far in the final proofing. It is anticipated that at the current rate a total of perhaps more than 200 unencoded forms will have been catalogued.

The missing forms are being added to the CDL database for tracking and future addition to the "Unihan Additions" database (as candidates for a future UTC submission to IRG), but it is recognized that:

(0) it is difficult to ascertain what is coming in the Ext C1 and C2 repertories, and so difficult to know what might need to be included in a future encoding submission (long-term tracking is necessary to follow-through in this);

(1) there is need for a systematic evaluation of this common PRC print source, and a

kXHC field might well be added to Unihan.txt, to aid in tracking unencoded characters;

(2) this issue highlights again the obvious need for some systematic treatment of the variant and simplification issue.

It is not expected that complete evaluation of this print source will reveal more than a few hundred unencoded characters. But, it is wholly unacceptable that, given the current size of Unihan's encoded character set, such problems still crop up, especially with regard to such a common dictionary in use in PRC for the last 26-or-so years. Clearly, simply adding more ideographs to the encoded UCS repertory is not going to correct such problems, as the set of simplified ideographs is more open-ended than the set of ideographs in general (which is itself very open-ended).

At the current rate, we fully expect that the set of simplified and variant forms as yet to be encoded will exceed the size of the currently encoded set of unified ideographs, in the near future (assuming that the repertories for future encoding can be developed). Following current practices, this process does not seem to be sustainable, and it does not seem that there is an adequate system in place for error-checking and maintenance.

A standard CDL-based variant mechanism (perhaps one such as that implemented by Wenlin in the CDL database, currently using an entire Private Use plane of variation selectors) would provide the most flexible longterm options for addressing this difficult problem.

Any use of variation selectors that that does not include use of CDL would be a short-sighted solution, inadequate for capturing the nuances of CJK ideographs, inadequate in the longterm for maintaining the integrity of UCS data.