L2/06-208

Additional Fields for the Unihan 5.0 database

John H. Jenkins
Apple Computer, Inc.
16 May 2006

In addition to the kCheungBauer and kCheungBauerIndex fields already approved by the UTC, I have recently found public sources for three additional new fields and would like UTC permission to add them to the 5.0 version of the Unihan database:

1) kFourCornerCode

The four-corner code(s) for the character. This data is partly derived from data provided by Hartmut Bohn, Urs App, and Christian Wittern available as part of the Unicode Han-Character Properties Database dated 23 July 1993. Additional data are derived from CCDICT 5.0.0, dated 2005.

The four-corner system assigns each character a four-digit code from 0 through 9. The digit is derived from the "shape" of the four corners of the character (upper-left, upper-right, lower-left, lower-right). An optional fifth digit can be used to further distinguish characters; the fifth digit is derived from the shape in the character's center or region immediately to the left of the fourth corner.

The four-corner system is now used only rarely. Full descriptions are available online, e.g., at <http://en.wikipedia.org/wiki/Four_corner_input>.

Values in this field consist of four decimal digits, optionally followed by a period and fifth digit for a five-digit form.

2) kWubi

The wubi (five-stroke) input code for the character. The core of these data are derived from the source files for mined-2000.11, an open-source Unicode text editor for Linux and Windows, with extensions from the SCIM sources.

The wubi input method divides a standard QWERTY keyboard into five "zones" for each of the five basic stroke shapes: the QWERT zone for falling-left strokes, the YUIOP zone for falling-right strokes, the ASDFG zone for horizontal strokes, the HJKLM zone for vertical strokes, and the XCBN zone for hooks. The Z key is available for use as a wildcard.

The user does a structural analysis of the character and, depending on the specific components involved, determines the wubi code.

Wubi codes are frequently used for input of simplified Chinese.

3) kHangul

The modern Korean pronunciation(s) for this character in hangul. The goal here is to shift data from the kKorean field, which has irregular romanization and other problems, to this field. This is also form used by ICU, which would make it easier for ICU to derive data for its own use from the Unihan database. (Indeed, the current dataset has been stolen ruthlessly from ICU.)

Depending on how Magda's time goes, we may have a fourth, as well:

4) kHanyuMandarin

This field contains the pinyin readings for ideographs in the *Hanyu Da Zidian*. The only entries are those explicitly in the *Hanyu Da Zidian* in pinyin. Pronunciations which can be deduced by the variant data in the *Hanyu Da Zidian* are not included, nor are pronunciations indicated in a non-pinyin system. The order of entries is that found within the *Hanyu Da Zidian* which is generally (but not always) that of modern frequency.

For a more accurate measure of modern Mandarin use, the kHanyuPinlu field should be used instead of kHanyuMandarin.

This field contains more data than the current kMandarin field and has a definite provenance, making it more useful in general than the kMandarin field is.

The field is fully populated and Magda is currently engaged in proofing it (and most of the way through, at that). Assuming we can get the proofing done in time and roll the corrections in, it would be nice to include this in 5.0.

Should Richard and Rick get the XHC data to me in time (action item 101-A62), I could also get that added.