L2/04-039

Proposed Changes to the Unihan Database

John H. Jenkins
Apple Computer, Inc.
28 January 2004

There are a number of changes I'd like to propose to the Unihan database in the Unicode 4.1 timeframe. If the UTC approves them, I'll go ahead and make them. I'm

1. Deprecate two fields, kAlternateMorohashi and kAlternateKangXi. Originally, the kMorohashi and kKangXi fields represented the official CJK-JRG/IRG indices for characters in the Morohashi and KangXi dictionaries. kAlternateMorohashi and kAlternateKangXi were intended for Unicode to use where our index values different from the official ones.

The IRG no longer actively maintains indices into Morohashi, but the kIRGDaiKanwaZiten holds the most recent official mappings, and the kMorohashi field the Unicode alternates. Similarly, the kIRGKangXi field now holds the official IRG mappings with kKangXi now holding Unicode alternates. The kAlternateMorohashi and kAlternateKangXi fields no longer serve any useful purpose and should be deprecated (or even dropped, if that's possible).

2. The kTang field holds a reconstructed seventh century pronunciation from *T'ang Poetic Vocabulary* by Hugh M. Stimson, Far Eastern Publications, Yale Univ. 1976. Because the data was originally in MacRoman, it was impossible to use three of Stimson's vowels: ɑ, U+0251; ɛ, U+025B; and ə, U+0259, so three different symbols available in MacRoman were substituted. Now that the Unihan database is stored entirely in UTF-8, there's no need to resort to kludges like this to represent these three vowels. I think we should switch to using Stimson's romanization as found in his book.

The kTang field is one of our incomplete fields, with only 121 entries. It would help in the process of filling in the field and using it were we to make the switch, and it wouldn't particularly inconvenience anybody since the field is of limited utility in its incomplete condition.

3. We maintain and are continually improving our mappings between simplified and traditional Chinese, but we don't have mappings for the Japanese national simplifications. I would like to add a kNationalSimplification field to hold this data. The syntax would be similar to the syntax for the kTraditionalVariant and kSimplifiedVariant fields, with the addition of a colon and a single letter (C, J, K, or V) after the Unicode code point.

We could therefore have lines such as:

```
U+5077   kNationalSimplification          U+5078:J
U+5078   kNationalSimplification          U+5077:C
```

This pair of entries would indicate that the two characters are Y-variants, but that U+5077 偷 is the form preferred in China and U+5078 偸 is the form preferred in Japan. The syntax is designed to allow similar data for Korea and Vietnam.

This field, if approved, would not be added to the public database until it is reasonably complete.

4. Cheung Kwan-hin and Robert S. Bauer of Hong Kong Polytechnic University recently published a monograph, *The Representation of Cantonese with Chinese Characters* in the *Journal of Chinese Linguistics*. This monograph is an excellent source of information on Cantonese as actually written and printed in Hong Kong. I've received permission from Cheung and Bauer to include data from their monograph in the Unihan database, and would like similar permission from the UTC. Again, the data would not be made public until it was complete.

Cheung and Bauer have two overall indices, one arranged by romanization (using the *jyut ping* system developed by the Linguistic Society of Hong Kong), and the other arranged by radical-stroke. The syntax for the kCheungBauer field would be to use both, thus:

```
U+34DF   kCheungBauer      pai1:018.05
```

This would make it possible to look the character up in either index. Alternatively, we could just use the standard page/position notation used with other dictionaries, with three digits for the page and two for the position.

For the record, Cheung and Bauer restrict themselves to Cantonese-specific ideographs and have well over a thousand entries. Unfortunately, they don't provide mappings to Unicode, only to Big Five and HK SCS. They also list approximately four hundred additional characters. My guess is that at least half of these are still unencoded, although I haven't checked any of them in Extension C yet.

5. I've defined an algorithm to generate a default sort key for Unihan characters. The algorithm is based on the Unicode code point and `kRSUnicode` field of the Unihan database and defines a 31-bit number which orders all the characters in Unihan by radical-stroke, and then by Unicode block (URO, Extension A, Extension B, Compatibility, Compatibility Extension), and then by code point. The algorithm allows for new characters to be added without altering the keys for existing characters.

The default sort key is already a field in the Unihan database. It's just one of the private fields we don't expose. I'm curious as to whether or not the UTC feels it would be useful to add to the public database.

6. One problem that we've got with our variant and pronunciation fields is that the authorities on which we base our data don't always agree. (This is the main reason why we've not filled in all of our variant fields.) Richard Cook has proposed an extension to the syntax we currently use for the Z-variant field which would also be handy for the other variant fields and the pronunciation fields.

Richard's extension would allow an entry in a variant/pronunciation field to be marked with two additional pieces of data: (a) the source, and (b) more detailed information on the type of variation.

For (a), the source is indicated by "=>" followed by one of our dictionaries.

For (b), the additional data is indicated by a colon followed by one or two letters. T (for *tóng* U+540C 同, used in dictionaries to mean *the same as*) would indicate that the source has marked the two characters as being identical in meaning. B (for *bù* U+4E0D 不, *not*) would indicate that the source has marked the characters as improperly used one for the other. A Z (for *zhèng* U+6B63 正, *correct*) would mark the form that the dictionary has indicated explicitly or implicitly is to be preferred.

To give an example, we could have:

```
U+292D8        kSemanticVariant   U+978B=>kHanYu:TZ
```

This would mean that U+292D8 𩋘 is a Y-variant of U+978B 鞋, according to the *Hanyu Da Zidian*. Morever, U+292D8 is defined (basically) with the words "同鞋", meaning they genuinely have the same meaning, whereas the definition of U+978B doesn't mention U+292D8 at all, which can be taken to mean that the preferred form is U+978B.

This extended syntax would be used in cases where one authority has ideosyncratic data not supported by other authorities, or in cases where two authorities might reasonably be expected to disagree. The former would be more useful in the pronunciation fields (e.g., Meyer-Wempe has an extraordinary number of unexpected Cantonese pronunciations), and the latter in the variant fields.

This extended syntax would be restricted to the two types of field. It would also have no impact on our basic typology for variants, viz.,

`kCompatibilityVariant` for compatibility variants as defined within Unicode;

`kTraditionalVariant/kSimplifiedVariant` for traditional and simplified Chinese variation;

`kZVariant` for forms which have unifiable shapes and identical meanings;

`kSemanticVariant` for forms which have nonunifiable shapes and identical meanings; and

`kSpecializedSemanticVariant` for forms which have nonunifiable shapes and overlapping meanings (e.g., the accounting numerals).

The one exception would be that the `kZVariant` field could now be used to mark pairs which are apparently unifiable and are confused for one another but strictly speaking have different meanings (e.g., pairs differing with *moon*/*meat* components).

7. The header for Unihan.txt is getting unwieldy. The only thing that's keeping it from being overwhelming is the

fact that it’s still such a small percentage of the file. I’m submitting to this UTC meeting a draft of a document entitled “A User’s Guide to the Unihan Database,” which I’d like to make into a UAX (or other appropriate document type). This document will hold an overall description of the database, how it’s maintained, what kinds of data it holds, and detailed information on the individual fields and their status. The header for Unihan.txt can then be reduced to a description of the file’s syntax, the standard legalese, and a pointer to the User’s Guide.

8. Inasmuch as Michel has added an IRG U-source to the WG2 data files, I’ve followed suit and added `kIRG_USource` (with twelve entries) to the Unihan database.