ISO/IEC JTC1/SC2/WG2/IRG N1861

Universal Multiple-Octet Coded Character Set
International Organization for Standardization

Doc Type: ISO/IEC JTC1/SC2/WG2/IRG
Title: Reading and Stroke-count Data in the Unihan Database
Source: John H. Jenkins, Unicode Consortium
Status: Liaison Contribution
Action: For consideration by the IRG
Date: 2012-06-12

Among the data contained in the Unihan database are two fields called kMandarin and kTotalStrokes. They are defined as follows:

kMandarin: The most customary pinyin reading for this character; that is, the reading most commonly used in modern text, with some preference given to readings most likely to be in sorted lists.…When there are two values, then the first is preferred for zh-Hans (CN) and the second is preferred for zh-Hant (TW). When there is only one value, it is appropriate for both.

kTotalStrokes: The total number of strokes in the character (including the radical), that is, the stroke count most commonly associated with the character in modern text using customary fonts.…When there are two values, then the first is preferred for zh-Hans (CN) and the second is preferred for zh-Hant (TW). When there is only one value, it is appropriate for both.

These two fields are used by the Common Locale Data Repository to provide default ordering for Chinese in various operating systems. Because these fields are seen directly by end users, it is vitally important that these fields be as accurate as possible.

Because stroke count data is maintained by the IRG and used in its own work, corrections to stroke count data are occasionally made. It would be very helpful if, when such corrections are made to already-encoded characters, the changes were individually recorded in a public IRG document, such as the Editorial Committee’s report. This would help the UTC provide the best data possible to end users.

The UTC also hopes that individual IRG members will continue to provide input to the UTC as it works to maintain and extend reading data for not only Mandarin but also Cantonese, Japanese, Korean, and Vietnamese. We thank IRG members for the help they have hitherto provided.