L2/03-197

Draft Document to Submit to WG2 on Simplified Chinese

John H. Jenkins

ISO/IEC 10646 has generally assigned (Chinese) simplified characters their own code points, separate from their traditional counterparts; e.g., U+8AAC (說) and U+8BF4(说). The decision to do so is based on a number of considerations, including the fact that the mapping between simplified and traditional forms is not always one-one, and the fact that the IRG G-source distinguishes characters in GB 12345-90 from their simplified counterparts in GB 2812-80.

The problem is that, while there are lists published by the PRC of official simplifications, most of these simplifications are obtained by applying a few general principles to specific cases. In particular, there is a set of radicals (such as U+2F94 KANGXI RADICAL SPEECH 言, U+2F99 KANGXI RADICAL SHELL 貝, U+2FA8 KANGXI RADICAL GATE 門, and U+2FC3 KANGXI RADICAL BIRD 鳥) for which simplifications exists (U+2EC8 CJK RADICAL C-SIMPLIFIED SPEECH 讠, U+2EC9 CJK RADICAL C-SIMPLIFIED SHELL 贝, U+2ED4 CJK RADICAL C-SIMPLIFIED GATE 门, and U+2EE6 CKJ RADICAL C-SIMPLIFIED BIRD 鸟). The basic technique for simplifying a character containing one of these radicals is to simply substitute the simplified radical. (Similar pairs exist for non-radical components, such as U+5340 區/U+533A 区.)

What this means is that at any time, any publisher of simplified Chinese text may create a new simplified form by merely simplifying the radical. There is, for example, the case of U+9D70 鵰, which is a kind of eagle. The "proper" way to write this character in simplified Chinese is to use U+96D5 雕, but there are instances in print of U+9D70 written with the simplified "bird" radical instead of the traditional one.

There is also the reverse problem, where the simplified form of a character is created (or encoded) first, and the traditional form derived (or encoded) later. Such a case is U+4882 轪/U+282E2 , a relatively recent character meaning "elevator," and used in Singapore and Hong Kong. As it happens, the simplified form (used in Singapore) was encoded as part of Extension A, and the traditional form (used in Hong Kong) encoded later as part of Extension B.

There are numerous instances in the current Extension C work being done by the IRG of characters which represent regular simplifications of existing traditional forms. There are also thousands of cases of traditional forms for which no simplification currently exists but which could potentially have one created by a publisher or font-designer at any time, given the productive nature of the Han ideographic script and simplification process.

The UTC notes that each case of a simplified/traditional pair encoded as such within ISO/IEC 10646 adds to the overhead of implementing Chinese support. Vendors must maintain increasingly large databases of such pairs for equivalence mappings. Font developers are also required to add new glyphs to fonts despite the fact that either the traditional or simplified form may naturally mesh with the overall font design, and despite the fact that both forms will never in practice occur within a single document.

Given this, the UTC recommends that WG2 in the future encode new simplifications of encoded

traditional forms (or vice versa) via the use of variation selectors, instead of the assignation of new code points. (This restriction would only apply to cases where a new simplified form has been created by application of the general simplification rules, and eliminate the problem of the relatively rare simplified characters which are simplifications of multiple traditional forms.)

This better reflects the productive nature of the script, simplifies font design, and makes normalization of Chinese text more straightforward. It would represent a departure from current policy and would leave implementors of ISO/IEC 10646 in the position of having two different solutions in place for different sets of simplified/traditional Chinese pairs, but this is no worse than the current situation of having some accented Latin letters available in precomposed form and others only via composition. On the whole, the UTC feels that the long-term benefits of using variation selectors for new simplified Chinese characters outweighs this awkwardness.