JTC1/SC2/WG2 N2476

Date: 8 May 2002

ISO/IEC JTC1/SC2/WG2

Source/Contributor Identifier: Unicode Techncial Committee

Title: **Variants and CJK Unified Ideographs in ISO/IEC 10646-1 and -2**

There are currently more than 75,000 CJK Unified Ideographs in ISO/IEC 10646-1 and -2. A large number of these are variants of one another. These variants come in different classes: pure z-variants, such as U+8AAA (說) and U+8AAC (説), where the glyphs involved are universally regarding just different ways of writing the same character and whose inclusion in ISO/IEC 10646 is largely for round-trip compatibility; simplified and traditional Chinese, such as U+8AAA (說)/U+8AAC (説) and U+8BF4 (**说**); accounting numerals, such as U+5341 (十) and U+62FE (**拾**); and so on.

Each class of variation has its own equivalence problems, but one fundamental problem is common to them all: end-users may want text to be treated is equivalent, even when variant characters are used. To give one instance which has been of some importance in early 2002, most users want simplified and traditional Chinese to be "the same" in internationalized domain names. Latin domain names, after all, are case insensitive. "Www.Unicode.Org" resolves to the same address as "www.unicode.org". There has been a strong push for "www.**同一碼**.org" (traditional Chinese) to resolve similarly to the same address as "www.**同一码**.org" (simplified Chinese). The inability to provide for this very nearly prevented Chinese from being used in internationalized domain names.

Programmers and users are being increasingly frustrated that as ISO/IEC 10646 becomes more pervasive, they are increasingly compelled to deal with a large number of variant characters some of which are only subtly different from each other and which cannot be automatically equated. Even within East Asia, most users are not aware of the subtle distinctions involved and can easily become confused when presented with a plethora of choices based on almost invisible differences. Some of the more egregious examples, as gathered by Rick McGowan, include:

Six variants for *feathered streamer*: U+7FFF 翿, U+26487 翇, U+2649B 翛, U+2649E 翞, U+264AB 翫, and U+264AF 翯.

Ten variants for *turtle*: U+4E80 **亀**, U+9F9C **龜**, U+9F9F **龟**, U+20074 𠒴, U+200FE 𠓾, U+24563 𤑣, U+27474 𧑴, U+2A6A7 龜, U+2A6A8 龝, and U+2A6BF 龜.

Thirteen variants for *business*: U+5546 **商**, U+2063E 𦏾, U+20E67 𠹧, U+20F83 𠾃, U+20FE7 𡃇, U+210A6 𡂖, U+210EC 𡃬, U+2111A 𡄚, U+2115F 𡅟, U+25AD0 𥖐, U+27DDE 賓, U+28757 鄗, and U+28DBC 鬮.

Fourteen variants for *disorder*: U+4E71 **乱**, U+4E82 **亂**, U+200F6 𠃶, U+200F9 **乹**, U+200FF 乳, U+209B8 率, U+209CE 率, U+209CF 率, U+22BA3 𢮣, U+24510 𤔐, U+24512 𤔒, U+24514 𤔔, U+24526 𤔦, and U+2452C 𤔬.

Twenty-seven variants for *slay*: U+6740 **杀**, U+6BBA **殺**, U+715E 煞, U+95B7 **鬁**, U+20112 𠄒, U+2205B 𢁛, U+2238A 𢎊, U+22393 𢎓, U+22F21 𢼡, U+22F22 𢼢, U+22F34 𢼴, U+22F45 𢽅, U+22F46 𢽆, U+22F58 𢽘, U+22F6F 𢽯, U+22F7A 𢽺, U+22F88 𢾈, U+22F8D 𢾍, U+22FD4 𢿔, U+22FF9 𢿹, U+2300E 𣀎, U+233C2 𣏂, U+23A86 𣚆, U+23A96 𣚖, U+23AA9 𣚩, U+2452E 𤔮, and U+2793F 𧼿.

One will quickly observe that many of these variants were added with Extension B.

It is vitally important that data be provided to allow developers, protocols, and other standards to deal with Han variants. This should not be taken to preclude individuals from providing more sophisticated handling; this can be something that can provide differentiation between products. It should also not be taken to mean that such data adequately defines means to interconvert between texts written using different sets of variants (particularly between simplified and traditional Chinese); this sort of process is too complicated and dependent on semantic analysis of the text to be made automatic everywhere.

What *is* needed, however, is something that allows at the least for a first-order approximation of equivalence. This would allow other standards and protocols, for example, to equate **同一碼** and 同一**码** or 說文 and 説文. It means, of course, that some false matches would also be possible; it would be up to the authors of the individual application, protocol, or standard to determine whether this were acceptable or not.

It is the feeling of the Unicode Technical Committee that there are two tasks needed to be undertaken to solve the Han

variants problem.

1) Data needs to be gathered on the existing variants within ISO/IEC 10646 and Unicode. This should include a classification scheme for the various categories of variant and include the various compatibility ideographs as well as the non-compatibility ideographs.

2) There are situations were some users, at least, wish to make a distinction between two variant forms for the same character within plain text. This is best handled using variant selectors. In the future, an ideograph should be separately encoded within ISO/IEC 10646 if and only if it is not a variant of an existing character. If it *is* a variant, then it should be represented using a variant selector.

Both of these have an impact on the work being done by the IRG. The UTC requests that the IRG be instructed:

A) To develop a classification scheme for variants within ISO/IEC 10646, allowing for different practices in different regions;

B) To gather data on existing variants within ISO/IEC 10646 classified in accordance with this scheme;

C) To require that variant data be provided as part of its work on Extension C, with the understanding that this data will be used to represent some characters submitted for Extension C via variant selectors. It should be made clear to IRG members that some of their characters will not be separately encoded in ISO/IEC 10646.