JTC1/SC2/WG2/IRG N914
Date: 2 May 2002

ISO/IEC JTC1/SC2/WG2/IRG
Ideographic Rapporteur Group
Secretariant: China

Source/Contributor Identifier:	Unicode Techncial Committee
Meeting:	For IRG #19 (Macao SAR, May 2002)
Title:	Variants within the CJK Unified Ideographs
Status:	Member body's contribution
Requested Action:	None; for information only

ISO/IEC 10646 contains over 75,000 ideographs. It should be no surprise that there are variant ideographs among them. These variants come in different classes: pure z-variants, such as U+8AAA (說) and U+8AAC (説); 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" to similarly resolve to the same address as "www.同一码.org". 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.

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.

Unicode 3.2 adds sixteen variant selectors. More will be added to Unicode and ISO/IEC 10646 in the future precisely to accomodate Han. The UTC feels that it is important to meet the needs of users by allowing for subtle distinctions between variant glyphs to be made in plain text. People who use an unusual form for a character in their name should be accomodated. Scholars dealing with an ancient text where multiple glyphs are used which may (or may not) have a distinct meaning should be accomodated. Modern texts which want to represent simplified Chinese characters not in official government lits should be accomodated. It is the intention of the UTC, however, that such accomodations should be made using varianat selectors. This permits people who need to make these distinctions to do so, but it does not burden everyone who supports CJK Unified Ideographs with having to track and directly incorporate the full body of equivalence data for Han variants.

Again, we must emphasize that this is a severe problem which is hindering the adoption of ISO/IEC 10646 or

limiting its ability to meet user's needs. If we desire to promote ISO/IEC 10646, it is important that we provide the data needed to solve it.

The UTC will ask WG2 in its Dublin meeting in May 2002 to instruct the IRG to do three things. The first is to develop a model for CJK Unified Ideograph variants which can handle the main classes of variant and allow for different useage in different locales. The second is to provide data on variants already found among the CJK Unified Ideographs in the standard. The third is to maintain with Extension C data variant information which can be used by WG2 to determine whether characters in Extension C should be represented by being separately encoded or by using a variant selector in connection with an already-encoded character.

The UTC thanks the IRG for the efforts it has already made along these lines, and in particular for requiring variant data be included with Extension C submissions. The IRG is without doubt the best body to provide a solution to this problem, with its unique combination of experience, knowledge of encodings, knowledge of East Asian text, and exemplary international cooperation.