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:	Japan's Proposal on Error Correction on Unified Ideographs in UCS
Status:	Member body's contribution
Reference:	IRG N903
Requested Action:	For discussion at IRG #19

The UTC agrees with Japan that it is a serious omission that there is no procedure for correcting errors in unification in the CJK Unified Ideographs found within ISO/IEC 10646. In this connection, the UTC would like to raise the following points.

1. There needs to be a more solid typology. The approach to handling a deunification depends strongly on what kind of characters have been unified. For example, a character from the KangXi dictionary which has been uncorrectly unified with a character from CNS 11643-1992 can more readily be deunified than one from JIS X 0213 which has been unified with one from CNS 11643-1992. An incorrect unification which involves rare characters can more easily be fixed than one which involves common characters.

Along these lines, although we agree with point one on pages 3 and 4 of Japan's paper, we feel that basing the disctinction on which character should be relocated solely on appearance may not always be appropriate. Again, if a character from the KangXi is incorrectly unified with one from CNS 11643-1992, then relocating the character from the KangXi has no impact on mapping tables and makes more sense, regardless of character shape.

2. The UTC feels that experts from WG2 and the UTC will more readily be able to discuss the issue of the IRG provides an exhaustive list of known deunifications which should be made. It would provide a better basis for understanding the nature of the problem and how it might be solved.

3. The UTC feels that it may be inappropriate to institute a set of general principles which cannot be overridden on a case-by-case basis. As we note in (1) above, there are different types of unification errors, and WG2 needs to retain the flexibility to handle them differently if it feels best.

4. Inasmuch as ISO/IEC 10646 is commony used in conversion and data normalization, particularly on the Web, it should not be forgotten that changing this mapping data—which is what a deunification would entail—has a real impact on data. Suppose, for example, I have a database with a Unicode front end and a JIS X 0213 back end. If the mapping data between the two is changed between the time data is entered and when it is retrieved, a user may be confronted with what they see changing without their permission. This sort of scenario is not at all uncommon, particularly as people are moving to use Unicode as a means of providing support for important character sets such as JIS X 0213 and collections such as HK SCS.

We should not let this prevent us from making corrections if such are needed. The UTC agrees completely with Japan that a solution is needed. Such corrections, however, need to be made with extreme caution.