

ISO/IEC JTC1/SC2/WG2/IRG

Ideographic Rapporteur Group

(IRG)

Source/Contribution Identifier : Ideograph Radical Ad Hoc  
Meeting : 20<sup>th</sup> IRG Meeting in Hanoi, Vietnam  
Title : Ideograph Unification  
Status : For use by the IRG

The IRG follows the unification rules given in ISO/IEC 10646, Annex S. The examples of the rules in Annex S are, however, typical examples, and not meant to be exhaustive. This means that there have been many cases where it wasn't clear whether two forms should be unified or not.

The Unification Ad Hoc group was requested to provide more detailed guidelines to the IRG on the issue of how to deal with doubtful cases.

The Ad Hoc group recommends that the IRG maintain a standing document listing the specific components which should be unified, and those which should not be unified. These lists should initially be populated with the examples from Annex S, and they should be updated as the IRG editors encounter more examples. As with Annex S, it should be understood that the examples in the standing document are not exhaustive.

For doubtful cases, the Ad Hoc group notes that the best thing to do would be to follow a course of action which could most easily be corrected if it proves to be wrong.

Where characters are from two different national or regional standards, then the mappings between those standards and ISO/IEC 10646 become a *normative* part of the standard. As such, neither incorrect unifications nor incorrect disunifications can be easily corrected. IRG editors should use the greatest possible care to make sure that no errors are made in unifying or disunifying such characters.

The situation is different for characters which are derived from other sources. Characters from dictionaries or informal and private character collections can be unified and disunified more easily.

For these characters, it seems clear that an incorrect *unification* can be fixed more easily than a false *disunification*. This is because although new characters can be *added* to ISO/IEC 10646, they can never actually be removed. Existing characters can be deprecated, which means that their use is strongly discouraged; but a character which once becomes a part of ISO/IEC 10646 is there forever.

Fixing an incorrect unification is not, however, without its own price. Some data will be encoded using the incorrectly unified character. Such data will have to be revised. Such a cost is minor, however, compared to the difficulties that will be caused by forever having the incorrect form available to confuse users and appear improperly in documents.

Because of this, the recommendation of the Ad Hoc group is that where there is doubt as to whether two forms should be unified or not, the null hypothesis is in favor of unification—that is, the editors should recommend unification in the absence of evidence to the contrary.

In the past, editors have tended to recommend disunification unless they were *sure* that characters should be unified. The recommendation of the Ad Hoc group is that editors now recommend unification unless they are *sure* that characters should be disunified.

The Ad Hoc group also recommends that a distinction be made between stroke normalization and derived simplifications for characters used in the People's Republic of China. As a rule, forms with stroke normalization should be unified, but derived simplifications should not (pending proof of their actually being used).