IRG N932

# ISO/IEC JTC 1/SC 2/WG 2/IRG
# Ideographic Rapporteur Group
# Secretariat: China

**Title: Problem with existing SuperCJK indexing system**

**Doc. Type:** National Body Contribution

**Source:** James Seng, Singapore

**Status:** to be discussed at IRG#20

**Date:** 22th September 2002

**Distribution:** ISO/IEC JTC 1/SC2/WG2/IRG

**Reference:** IRG N907, N928

The SuperCJK document sorts the ideographs according to the KangXi index, which is derived from the "Radical", "Stroke Count" and the "First Stroke" of the ideograph.

A proposed ideograph has to provide its KangXi together with its "Radical", "Stroke Count" and the "First Stroke" which we will match against the SuperCJK for duplicate. Therefore, the tuple (Radical, Stroke Count, First Stroke) form the basic key for indexing ideographs.

Currently, there are several problems with this index system:

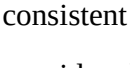
1. **Ambiguous Radical**

   Sometimes an ideograph has more then one possible Radical. For example, IRG N928-00392  , the listed radical is  although either  and  are also valid radical.

   Sometimes an ideograph may not have any possible Radical. For example, IRG N928-00457  , the radical is undeterminable. In such case, we would "by convention" use the first stroke as the radical, although that is also not consistent.

2. **Ambiguous Stroke Count**

   Due to the differences in writing style across region, simplification process and others, we have ideograph that is not consistent in their stroke count. For example, the radical  could either be 4 stroke as in  or 3 strokes as in .

By convention, we would use the stroke count as defined in KangXi but that is not consistent and sometimes not feasible to do so. Otherwise, 月 could have to be considered as 6 stroke as in 肉, not 4 stroke as we would normally count. Likewise,  and  would have counted as 6 not 3 or 4.

**3. Ambiguous First Stroke**

Once again, due to the differences in writing style across region, the first stroke may be different. For example, for the ideograph 女, at least in Chinese writing style, would have its first stroke as 5 but some country have the first stroke *very* consistently as 1.

There are also numerous occasion when a ideograph similar to 向 have its first stroke either as 3 or 4. (The “correct” answer would be 3.)

Indexing system should be deterministic and unambiguous. But with these problems, the existing indexing system is far from ideal. The current workaround is to do a “fuzzy” match, considering the other possible radicals, other possible stroke count and other possible first stroke. Such workaround results in inaccuracy and inconsistency and therefore, defends the purpose of having an index in the first place.

While the KangXi indexing system has served us in the past, it is apparently we need a better indexing system now. The IRG have to solve this indexing problem quickly before the task of doing “fuzzy” match become impossible.

We propose that an adhoc group to be form in IRG #20 to discuss this problem and to makes it recommendation to the group.

--- End ---