Universal Multiple-Octet Coded Character Set
UCS

ISO/IEC JTC1/SC2/WG2 IRG N 1173R

Date: 2005-12-01

Doc. Type:	Member body contribution
Title:	Comments on IRG N1153
Source:	TCA and Unicode consortium
Status:	Input to IRG
Action:	To be discussed
Distribution:	IRG Members and Ideographic Experts
Reference:	
No. of pages:	
Medium:	Electronic

TCA and Unicode believe that the use of IDS may greatly help CJK standardization work. But we also believe that the method may cause some unexpected problems. The following are some of our concerns:

1. Is component position and ordering information essential? Is it really helpful to the process of identifying duplicates?

   (1). **High training cost**. To use this method to decompose characters, people need to spend much time learning and getting used to the rules. As is well known, there is variation in the writing order of characters and in identification of components, and it is not easy to force users of IDS to change their habits, especially since people will be working quickly, in order to accomplish the work in a reasonable amount of time.

   (2). **Human error**. Even following the guidelines, people may decompose the same character in different ways. As was seen in yesterday's review, even a well trained person can easily decompose a character incorrectly. More complicated rules mean more human effort and more human error.

2. The main purpose of character decomposition is to identify possible duplicate characters. Why then do we not we just decompose the characters in the most simple way possible? The following are some suggestions and comments on the document IRGN1153.

   (1) **DO NOT restrict decomposition order/direction**. Multiple decompositions might be permitted. For the purpose of character comparison, relative component position and order do not provide sufficient additional information, to compensate for the added complications which they introduce. Normally two different characters contain different

components. We can simply compare two characters without any radical position or order information. Even if the comparison system points out two different characters as possible duplicates, human verification is still required, and it is easier for the human eye to compare simplified descriptions than more complicated ones. The savings in time and effort would be significant.

(2) **DO NOT restrict the depth of component analysis**. For example, the character "彬"might be decomposed in any one of the three following ways: "林+彡", "木+杉" or "木+木+彡". If "彬" is decomposed as "林+彡" then the comparison system should automatically decompose "林+彡" into an undecomposable component level (in this case is "木+木+彡") prior to doing the comparison.

Comment: Where is the "existent radical data"? When we start to decompose the characters, we are building up the mapping data.

3. **Conclusion**:

The purpose of using character decomposition is to reduce the workload and also the human mistakes. But if the decomposition rules are overly complicated, the learning process will be too long, the results themselves will contain too many errors, and the purpose of using the system will be defeated.

Appendix.

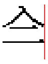
The following examples show how easy it is to make "mistakes".

1-2器→⿳吅犬吅

WHY NOT?⿱哭吅

Comment: Unicode:5405

4-7] →⿶凵⿱人二

WHY NOT?⿶凵

Comment: Unicode:204DE

4-7 凶→⿶ 凵乂

WHY NOT?⿶ 乂凵

4-9]厚→⿸厂⿱日子

WHY NOT?⿸厂㫗

Comment: Unicode:3AD7

4-10]貳→⿹弋⿱二貝

WHY NOT?⿹弍貝

Comment: Unicode:5F0D