ISO/IEC JTC 1/SC2/WG2/ Ideographic Rapporteur Group

Source / Contribution Identifier: MS
Meeting: IRG#25 @ Berkeley, CA
Title: Responses to IRG N1173, Comments to IDS Guidelines
Status: Personal Contribution
Short Description: to be discussed on IRG #25
Proposed Conclusion / Requested Action: to be discussed at IRG #25
References: IRG N1153, N1154, N1173

1. Background

The purpose of this paper is to remove conflicts between N1153 and N1173. Since N1153 and N1173 are both in favor of using IDS for IRG work, the difference of the two documents is on technical details.

Most of the points stated in N1173 discuss some trade-offs. The author considers they have a reason. However, we have no data to measure pros and cons between alternatives, it is somewhat hard to find the optimum setup at this moment. It is important that IRG to work as a group under the same understanding, The author believes that Japan can agree on any modification to the document IRG N1153 or even on simply discarding it at all, if the IRG concensus supports the idea.

2. Brief comments on the points in N1173

In short, the document N1173 proposes to remove some or all of *guidelines* over the UCS definition of IDS. As stated in N1153, the guidelines are set just to give a common ground and reduce the difficulty of machine detection of ideographs to be unified. Japan already has a prototype implementation of the IDS comparison tool, and the algorithm used in the program takes care of most of the typical IDS differences from a same ideograph. (See IRG N1154 for the details of the algorith the current version of the program is based on.) So, in most cases, we don't need the guidelines described in N1153. The N1173 is correct on the point.

However, there are some exceptional cases the current algorithm cannot handle well. The intended purpose of the guideline was to minimize such cases, and to maximize the usefulness of IDS. On the other hand, as stated in N1173, following the guidelines in N1153 requires additional effort by IRG editors. Since the case the current comparison program cannot handle well is relatively rare, the ratio the program fails to detect

duplicates will not greatly change if we don't follow guidelines. Discarding some or all of the guidelins may be, hence, a good idea, when considering the overall workload of editors.

3. Some details

3.1 Regarding the point 1.

I personally agree that the component position and ordering information are not essential, although I believe they are helpful.

For example, in Annex S unification rule, two ideographs consists of ideontical set of components but with different relative positions are considered not to be unified, position and ordering may be helpful to reduce the number of possibly-to-be-unified pairs. (See Annex S.1.4.2.)

However, the typical cases like examples listed in Annex S.1.4.2 are automatically handled by the IDS mechanism itself, and we don't need additional guidelines as in N1153 to cover the cases.

3.2 Regarding the point 2.

As stated in 2 above, the current program can handle the said cases. For example, it finds three IDSs, "□□林 彡", "□□木杉" and "□□木木 彡" possiblly identical. We don't need any additional guidelines as in N1153 to handle this case.

3.3 Regarding some of the examples listed in Appendix

4-7 凶 → □ 凵乂 WHY NOT?□ 乂凵

Because the standard says so. This *restriction* is by 10646, not by N1153. See Annex F.3.2 and Table F.1. The author simply wants to follow the standard.

For other "WHY NOT"s in the appendix, the author has no objection; yes, they can be.

4. Conclusion

The purpose of the guidelines in N1153 is to provide a common ground on the use of IDS, to reduce the difficulty of the machine-matching, and ultimately to make editors' time-consuming jobs easier. Since they are not essential, we can remove the guidelines without seriously limiting the usufullness of IDS in Extension C reviewing tasks.

The author believes Japan has no objection not to use the guidelines described in N1153.

The author still hopes IRG to work collaboratively.

(END OF DOCUMENT)