ISO/IEC JTC1/SC2/WG2/IRG N1382

UTC Recommendations for IRG Principles and Procedures

John H. Jenkins

14 November 2007

I've been giving some thought to IRG principles and procedures and have a couple of recommendations I'd like to make.

There are two specific areas I'd like to focus on: the IRG's *collective memory*, and the *amount of data available* on individual characters.

One of the features of the original work done by the IRG's predecessor, the CJK/JRG, was that it was reproducible. That is, given the same sources and the same principles, another group of people would (in theory) be able to do the same work and reach pretty much the same outcome. This was largely because the IRG's work was based on widely available standards and dictionaries.

As the sources for the IRG's work become more obscure, this aspect of our work is being lost. Not only is it unlikely that a second group of people could start out with the same data and end up with the same results, but it's very unlikely that even the IRG itself would end up with the same results. This is because we don't track the decisions we make—particularly the editorial decisions—in any readily available form.

This also means that the editorial committee will have to periodically revisit unification issues it has already decided. There are cases where we found component shapes which are sometimes unified and sometimes not. We can generally tell what decision was made in the past, but we can't tell why it was made.

I feel that this problem can best and most easily be resolved by the editorial committee making its minutes from past meetings available at least to IRG members. We would probably not want these to be documents available on the IRG's public web site, but they should nonetheless be made available for editorial committee members.

If editorial committee members had access to the minutes from past meetings, we could more easily determine what precedents to apply and save time currently spent recreating decisions that were made in the past.

The other issue has to do with the amount of data that IRG members are making available regarding the characters in their collections. Certainly this is true for the UTC characters. For most of the characters we have submitted to the IRG, we have specific source references as well as pronunciation data and definitions. This data is

currently available only internally to the UTC itself (and not even always there).

The IRG is already expecting members to provide more information about submitted characters. I think tendency is a good one and should be encouraged. What would be most useful would be for this data to be available to editorial committee members in a uniform way, or at least in documents which are maintained in a database or on a Web site such as the IRG's public site.

My point here is not so much that the IRG needs to require more data from member bodies, but that it needs to make sure there is a known, systematic way of retrieving that data in its current form after it has been received. This could be done as simply as having a special page on the IRG Web site where the initial proposals and the evidence documents are kept accessible. These documents could be kept updated and current by the various IRG members.

It would also be helpful if evidence documents were divided into documents with text-only data (pronunciation and defi nition data) and documents with scans of sources. This would make it easier to look up data in the middle of a meeting, because we wouldn't have to wait to search for text inside a PDF when it's only text we want to find.