L2/03-449

Report on IRG Meeting 21

Guilin, China, 17-21 November 2003

John H. Jenkins

IRG meeting number 21 was held in Guilin, China, from 17-21 November 2003. The US/Unicode representatives were John H. Jenkins (Apple), Tom Bishop (Wenlin), and Hiura Hideki (Sun). All IRG members were in attendance except for Singapore and North Korea.

Copies of *The Unicode Standard, Version 4.0* were distributed to IRG delegations at the meeting, as well as to the editors and the maintainers of the Web site. The Hong Kong SAR delegation generously helped get the books to the meeting.

The key documents from the meeting have been submitted as UTC/L2 documents: The meeting resolutions, the report from the Editorial Subcommittee, the report from the Subset Subcommittee, and the informal report on the encoding of old hanzi.

IRG members reported on their status at the beginning of the meeting. Of particular note were the following announcements:

The Hong Kong SAR is planning a revision of HK SCS to be released in late 2003 or early 2004. Two new unique ideographs are included in this revision. The Hong Kong SAR delegation is firmly committed to having all HK SCS ideographs mapped to non-PUA areas of Unicode and will include these two new ideographs in their next submission.

Taiwan will be issuing a new version of CNS. There are numerous new ideographs in this character set, with no indication as yet as to how they will be mapped to Unicode.

South Korea made a formal withdrawal of 3500 ideographs from their Extension C1 proposal. As they did not indicate which 3500 ideographs they were withdrawing until the meeting was underway, the work on Extension C1 remained stalled. The editorial subcommittee worked on an aggressive new schedule which, it is hoped, will get it back on track. If all goes well, the IRG may actually have a version of Extension C1 in time for the June WG2 meeting.

The Macao SAR is creating a "Macao Common Chinese Character Set" derived from newspaper use, and Macanese proper nouns (both place and personal names). They currently have 4954 characters; frequency data is included. This is intended as input for the IRG's subset work.

The bulk of the meeting was spent divided into three groups: the Editorial Subcommittee, the subset subcommittee, and an "old hanzi" ad hoc. The last-named dealt with issues involved with the encoding of older forms of East Asian ideographs. I attended and co-chaired the editorial subcommittee; Hideki attended the Subset Subcommittee, and Tom was a member of the Old Hanzi Ad Hoc.

The Editorial Subcommittee spent the bulk of its time on glyph or unification errors. There's nothing that can be done about the latter at this point, beyond pointing them out and using them as reminders of the existing unification rules. The Editorial Subcommittee did, however, resolve to keep a text file listing the cumulative borderline cases which it considered, to aid in the group memory and avoid having to reconsider decisions without knowing precisely how they were decided in the first place. It also adopted a sort of inverse to the formal non-cognate rule: in the case of a doubtful unification, if both forms occur in one of the IRG's standard

dictionaries and are explicitly given the same pronunciation and meaning, they are to be unified.

The Subset Subcommittee tried to come up with a formal subset which would meet the most important day-to-day needs of IRG members and which would be available as quickly as possible. Unfortunately, the third goal, of keeping the subset under 10,000 characters, proved inconsistent with the other two, particularly the second. The final count is on the order of 10,600 characters. (I suggested that at this point we may as well make it exactly 10,646 characters in size, but that suggestion seems not to have been treated seriously.)

The subset is currently dubbed IICore. The current version of IICore (1.1) is available on the IRG web site, <http://www.cse.cuhk.edu.hk/~irg/irg/IRG21.htm>. The IRG has requested feedback on IICore from the next UTC meeting.

The Old Hanzi Ad Hoc represents the first step towards dealing with ancient forms of Chinese. The Ad Hoc concluded that the various forms of old hanzi should be encoded in their own blocks, separate from the main ideographic blocks and using slightly different unification rules. This will also need to be discussed at the next UTC. Mr. Zhang will work with Mike Ksar to see what formal process needs to be undertaken in order to actually start on the work of old hanzi encoding.