L2/08-264

Title: *Proposal to add SuperCJK 14.0 index data to Unihan*
Authors: Richard Cook, Ken Lunde
Date: 28 July 2008

This is a proposal to add 70,205 records of indexing data to Unihan, for inclusion in a future release.

"SuperCJK_14.0_with_index.pdf" (IRG N802) was the final version of the multi-column CJK chart resulting from original IRG Extension B production work. Created in July 2001, it includes (at most four) large representative glyphs for each of 70,205 CJK characters (~ 111 MB):

```
http://www.cse.cuhk.edu.hk/~irg/irg/CJK/SuperCJK140_IRGN802.zip
```

This 2,128 page document remains a key point of reference for IRG and for developers, however the size and (radical/stroke) organization of the file severely limit accessibility.

To improve access, we have generated index data identifying the position of each character, and have validated this data against data independently generated by other IRG delegates (Japan NB).

Each line of the accompanying datafile

```
supercjk14-index-col-pos.txt
```

has the following form

```
03400    0002.1.07
```

mapping the Unicode code point to the *page.column.row* in *SuperCJK 14.0*.

We propose that these mappings be added to Unihan and given *informative* or *provisional* status.

The character coverage breaks down as follows:

CJK Unified Ideographs URO:	20,902
CJK Unified Ideographs Extension A:	6,582
CJK Unified Ideographs Extension B:	42,711
CJK Compatibility Ideographs:	10
Total	**70,205**

Note that the ten CJK Compatibility Ideographs are among those twelve considered CJK Unified Ideographs; two are thus missing (it is not clear if IRG was ever aware of this).

```
U+FA1F
U+FA23
```