L2/11/020

Title:	**Request for changes to the Unihan kRSUnicode field**
Source:	**Andrew West**
Status:	**Individual Contribution**
Action:	**For consideration by UTC**
Date:	**2011-01-24**

---

1. Introduction

ISO/IEC 10646 3rd Edition, which is currently under ballot, now incorporates a "Radical Stroke index" field in the "Source references for CJK Unified Ideographs" file (see <http://www.itscj.ipsj.or.jp/sc2/open/02n4168/CJKU_SR.txt>) and the "Source references for CJK Compatibility Ideographs" file (see <http://www.itscj.ipsj.or.jp/sc2/open/02n4168/CJKC_SR.txt>). This field corresponds to the kRSUnicode field in the Unihan_RadicalStrokeCounts.txt file of the Unihan database. However, there are some discrepancies between these two fields which should be addressed so that users using the CJKU_SR.txt and CJKC_SR.txt files get the same results as users using the Unihan_RadicalStrokeCounts.txt file.

2. Multiple Radical/Stroke Counts for 32 Characters

The CJKU_SR.txt and CJKC_SR.txt field have exactly one radical/stroke index for each character, whereas the Unihan kRSUnicode field may have one or two radical/stroke counts for each character (Unicode Standard Annex #38 allows for an unlimited number of radical/stroke counts for each character, but at present no character has more than two radical/stroke counts). In fact, only 32 out of the 75,394 characters in the Unihan database have a kRSUnicode entry with two radical/stroke counts (see appended table).

There are two main use cases for the kRSUnicode field: a) to help construct radical/stroke lookup tables for Han ideographs; b) to use as a sort key for sorting Han ideographs by radical/stroke counts. Note that kRSUnicode is the only Radical/Stroke field that is applied to all characters in the Unihan database, and so it is the only field likely to be used for these purposes. The provision of two radical/stroke counts for just 32 out of 75,394 characters does not really help for the first case, as there are many thousands of characters which are classified under different radicals or with different stroke counts in different dictionaries, and so the data in Unihan is hugely incomplete if this is the intention for providing multiple radical/stroke counts. On the other hand, provision of multiple radical/stroke counts is extremely unhelpful for users who want to sort Han ideographs by radical/stroke counts using the kRSUnicode field, as they cannot be sure which of the two radical/stroke counts provided for these 32 characters should be used as its sort key, with the result that different implementations may result in different sort orders with regard to these 32 characters, which is extremely unsatisfactory. Providing a single nominal radical/stroke index for each Han ideograph, consistent between the Unicode and ISO/IEC 10646 standards, seems to be the most satisfactory solution. If alternate Unicode radical/stroke counts are deemed necessary, then these should be provided in a separate, new field (e.g. named kRSUnicodeExtra).

Therefore, I would like to request the UTC to:

1. Remove one of the two radical/stroke counts provided for the 32 characters with two radical/stroke counts under the kRSUnicode field;
2. Change the definition of the kRSUnicode field given in Unicode Standard Annex #38 "Unicode Han Database (Unihan)" to allow only one radical/stroke count for each Han ideograph;
3. Ensure that the kRSUnicode field for each Han ideograph is the same as the Radical Stroke index in CJKU_SR.txt and CJKC_SR.txt, and that any future changes to kRSUnicode are reflected in CJKU_SR.txt and CJKC_SR.txt.

3. Different Radical/Stroke Counts for 2 Characters

Two characters (U+2305A and U+2AAA9) have different radical/stroke counts in CJKU_SR.txt compared with the Unihan database (see appended table). In the case of U+2305A, "67.8" given in CJKU_SR.txt seems to be correctish ("67.7" better reflects the glyph shape of the character in the code charts, and is consistent with the stroke count given for other characters with this component), and "66.11" given in Unihan seems to be erroneous. In the case of U+2AAA9, both "53.15" (CJKU_SR.txt) and 55.15 (Unihan) are valid alternatives, but for the sake of consistency with the radical of its surrounding characters, I recommend assigning "53.15" as its kRSUnicode value.

4. Discrepancies between Unihan and CJKU_SR.txt and CJKC_SR.txt

Character	CJKU_SR/CJKC_SR	kRSUnicode
U+3687	35.6	35.6 66.6
U+4347	121.5	121.5 121.4
U+49F9	172.5	172.5 53.10
U+4E2C	90.0	90.0 90'.0
U+521D	18.5	18.5 145.2
U+52D7	72.7	72.7 73.7
U+5364	25.5	25.5 197'.0
U+553E	30.8	30.8 30.9
U+56CA	30.19	30.19 145.17
U+5782	32.5	32.5 33.5
U+6376	64.8	64.8 64.9
U+6B3D	76.8	76.8 167.4
U+6C3D	11.4	11.4 85.2
U+6E20	85.9	85.9 75.8
U+7262	93.3	93.3 40.4
U+7740	109.7	109.7 123.5
U+8006	125.4	125.4 72.6
U+8FB6	162.0	162.0 162'.0
U+8FF8	162.6	162.6 162.8
U+9077	162.12	162.12 162.11
U+9756	174.5	174.5 117.8
U+9EC4	201.0	201.0 201'.0

Character	CJKU_SR/CJKC_SR	kRSUnicode
U+9F50	210'.0	210'.0 67.2
U+9FBD	86.11	86.11 195.4
U+9FC2	159.11	159.11 196.7
U+FAAA	109.7	109.7 123.5
U+FAC8	174.5	174.5 117.8
U+221A1	5.13	5.13 39.11
U+22605	61.5	61.5 110.4
U+2305A	67.8	66.11
U+243F2	86.12	86.12 86.11
U+24B96	98.8	98.8 98.10
U+2AAA9	53.15	55.15
U+2F835	27.4	27.4 86.2