L2/16-223R

Re: **Augmenting data for CJK Radicals and Strokes**
From: **Mark Davis, Ken Lunde**
Date: **2016-07-28**
Drafts: **https://goo.gl/oHFYwy, https://goo.gl/VQZvvr**

To analyse Unified Ideographs in terms of ideographic description sequences (IDS), those IDS are defined in terms of "Ideographic | Radical | CJK_Stroke" (TUS section 18.2). However, when you start to use the Radicals and CJK_Strokes, you run up against a lack of data about them. There is a mapping from certain of the radicals to strokes and Unified Ideographs in CJKRadicals.txt, but it is incomplete, and also sometimes inconsistent with the usage in the UCA and NamesList.txt. There are also other uses for more complete data, besides IDS.

Proposal

Produce a PRI that as a part of Unicode 10.0, we include Unihan-like metadata where possible for each of the CJK radicals and strokes (that is: CJK Strokes, CJK Radicals Supplement, Kangxi Radicals).

A. Include in the UCD a data file that is more complete than CJKRadicals.txt.
B. Create new character properties reflecting that data.
C. Use the data to update NamesList.txt and the UCA data, if and where appropriate.
    a. Supplement the NamesList with addition mappings where the analysis identifies missing or appropriate cross references.
    b. Review the UCA weight assignments of Radicals to find the most appropriate Unified Ideographs to equate them to.

The proposed character properties would be, for each CJK radical or stroke:

- **RadicalStrokeCount** (RSC): the number of strokes in the radical or stroke
- **RadicalNumber** (RNU): the radical number for the character (using the ' convention for simplified), if any
- **RadicalUnifiedIdeograph** (RUI): the best corresponding CJK Unified Ideograph(s), if any

This is not a proposal to replace CJKRadicals.txt, which some people need in the current format, but rather to provide additional information. The above are draft property names and can be changed, of course.

Problems

- The CJKRadicals.txt file doesn't cover all of the Radicals, and is really serving a different purpose.
- Some of the information on radicals is in the NamesList and UCA, but not in an easily machine-readable form.
- The UCA and NamesList mappings are incomplete for the CJK Radicals
- CJKRadicals.txt does not provide stroke data.
- The CJKRadicals.txt file is also in a somewhat awkward format. Unlike other UCD files, the source character is not the first field, so we don't treat it as defining a property, and the data thus isn't normally available as UCD properties.

Preliminary Analysis

The attached data file (https://goo.gl/VQZvvr) contains an initial cut at supplying the missing data, using some processing from multiple sources. *It is not final form: it needs some fleshing out and adjustment as described below.*

Here is an example from the file.

2F54;	4;	85;	6C34;	#	水	→	水	KANGXI RADICAL WATER;	CJKRadicals.txt, NFKC
2EA1;	3;	85;	6C35;	#	氵	→	氵	CJK RADICAL WATER ONE;	UCA, NamesList
2EA2;	5;	85;	6C3A;	#	氺	→	氺	CJK RADICAL WATER TWO;	UCA, NamesList
2F54;	4;	85;	23C71;	#	水	≠	▯	KANGXI RADICAL WATER;	kRSUnicode

For the data lines, the following fields are used:

Field 0:	The source code point (radical/stroke)
Field 2:	**RSC**: the number of strokes in F0. This is obtained by looking up total strokes of the target Unified Ideograph (F3)
Field 1:	**RNU**: the radical number (with '). This is obtained from CJKRadicals.txt where possible, and otherwise from http://www.wikiwand.com/de/Unicodeblock_CJK-Radikale,_Ergänzung
Field 3:	**RUI**: the target Unified Ideograph(s). This is obtained from various "causes", listed in the comments after ';'. These causes are:
	• CJKRadicals.txt
	• NFKC
	• UCA
	• NamesList
	• kRSUnicode
	• kRSAdobe *(abbreviation for kRSAdobe_Japan1_6)*
Comment:	(source => target) <name> ; <causes>

The kRSUnicode and kRSAdobe data are used to get a target Unified Ideograph by searching for those with a radical, and zero remaining strokes. In such cases, the Unified Ideograph represents the radical alone. The source character (F1) has not been selected from the available radicals (pending agreement by experts); instead the line is commented with ? for the source code point and its name.

This is not final form: it needs some fleshing out and adjustment, and recasting as a normal property file. In particular, the most appropriate CJK Unified Ideograph would be picked for the radical or stroke. Only a small number of cases would be worthy of multiple values. For example, U+2F4A (Radical #75) would have only the following set of data in the data file

2F4A ; RSC ; 4 ;
2F4A ; RNU ; 75 ;
2F4A ; RUI ; 6728 ;

This would be done by picking the most appropriate CJK Unified Ideograph for the radical or stroke. Only a small number of cases would be worthy of multiple values, such as U+2FCA (Radical #203):

2FCA ; RSC ; 12 ;
2FCA ; RNU ; 203;
2FCA ; RUI ; 9ED1 | 9ED2;

The draft data includes characters in the fourth column that have zero (or less) residual strokes, but that is only for initial comparison; it would not appear in the final data.

The draft data has final status column with values “keep”, “drop?”, and “drop”, indicating the current status. These are also color-coded for visibility.

There is one outlier among the radicals, one that doesn't have a target Unified Ideograph.

2E80;	0;	7;	;	#	▫	→		CJK RADICAL REPEAT;	

There are also some cases where the "main" mapping data appears suboptimal, at least based on my fonts.
For example, it appears that the best target Unified Ideograph for ⻊ CJK RADICAL FOOT would be ⻊ (27FB7), not 足 (8DB3).

2F9C;	7;	157;	8DB3;	#	足	→	足	KANGXI RADICAL FOOT;	CJKRadicals.txt, NFKC

2ECA;	7;	157;	8DB3;	#	⻊	→	足	CJK RADICAL FOOT;	NamesList
2ECA;	7;	157;	27FB7;	#	⻊	→	⻊	CJK RADICAL FOOT;	kRSUnicode

If that is the case, then the NamesList mapping appears to be in in error; the better mapping would be to 27FB7 (that may not have existed at the time the NameList information was composed). Any such choices need expert review, of course.

The CJK Stroke characters don't have such sources of data, but it would also be useful to supply the same sort of information for them, *where possible*. From n3063, we have the following information. We should be able to use this data and the IDS data to provide a stroke count and in some cases a target Unified Ideograph character, in the same format as above, then review by experts. This data is found in the second sheet.