Subject: Proposed change of 4 characters in UCA
From: Mark Davis ☺ <mark@macchiato.com>
Date: Fri, 30 Jul 2010 13:33:00 -0700
To: l2doc@unicode.org, utc-chair@unicode.org

L2/10-275

Please add to the registry and agenda.

=========

1. There are only a couple of instances where currency characters are not grouped together, or where punctuation marks are not below Latin (see http://unicode.org/charts/collation/) I propose that we correct this in 6.0.

U+20A8 (₨) RUPEE SIGN
U+FDFC (﷼) RIAL SIGN
U+19DE (᧞) NEW TAI LUE SIGN LAE
U+19DF (᧟) NEW TAI LUE SIGN LAEV

(These are sorting with the respective scripts, rather than with similar symbols and punctuation.)

2. Grouping Punctuation.

For discussion, but no action in 6.0.

Longer term, it would be useful to consider grouping the punctuation together in Variable, below most symbols (except perhaps a few punctuation-like symbols).

Punctuation and Symbols generally want to be considered differently in collation. Collation sequences are closely coordinated with searching, and while it is quite common to ignore spaces or (),;..., people don't normally want to identify "INY" with "I♥NY" in searching or sorting. Because they are currently intermixed in the Variable category, this is not really feasible.

For collation reordering, this is also important, because we would like to be able to let people reorder scripts and other semantically important classes in collation (eg Cyrillic before Latin). The following categories are those important classes: punctuation, decimal numbers, Sc, and other symbols. For example, DIN 5007 and other standards specify that numbers should sort after letters. These categories do not have to be completely "pure", but should basically contain all the characters in that "class", with perhaps some intermixed characters that behave similarly. But having them be contiguous allows reordering to treat them as a reorderable chunk.