L2/15-073

Subject: **Progress on UTS #39**
From: **Mark Davis**
Date: **2015-02-03**
Live link: **link**

I had the following action.

> **[140-A76] Action Item for** Mark Davis, Editorial Committee: Issue a proposed update of *Unicode Technical Standard #39, Unicode Security Mechanisms* for V8.0.

Implicit in that action (though unfortunately not stated) was to also do the consensus 140-C21.

> **[140-C21] Consensus:** After releasing *Unicode Technical Standard #39, Unicode Security Mechanisms* for V7.0, post a proposed update of *Unicode Technical Standard #39, Unicode Security Mechanisms* for version 8.0 that removes the SL, SA, and ML tables, and documents how they can be derived instead, and makes the appropriate changes in the text.

I've completed a first pass at the text and data:

- text: http://www.unicode.org/repos/draft/trunk/reports/tr39/tr39.html (but still needing editorial review)
- data: http://www.unicode.org/repos/unicodetools/trunk/unicodetools/data/security/8.0.0/

For the data, I've removed the SL, SA, and ML tables. Remaining tasks to do are:

1. Revert the target characters to using the same algorithm as before 7.0 (so that it again favors ASCII over obscure symbols.
2. Incorporate other changes as per UTC actions, plus other feedback from the public.
3. Post within a month, so that we can take in any feedback before the May meeting.

Recommended for 8.0

There are also two items that I recommend we do for 8.0. Both of these involved very small tooling and text changes.

1. The **Type** values in http://www.unicode.org/reports/tr39/#Identifier_Modification_Key are primarily informative. We currently derive most of historic and limited-use from UAX #31, but the terms are not the same. I recommend that we align them to help make the tables and the derivation clearer, for example, by renaming the **Type** values "historic" to "exclusion", and splitting "limited-use" into "limited-use" and "aspirational".
2. For consistency, I recommend that we use formats for Idmod Status and Type values that follow the identifier syntax and style of the UCD, eg **Limited_Use** instead of **limited-use**. This also makes them easier to use as programmatic identifiers.

Recommended for 9.0 draft

For 9.0, I think we should have an action to investigate the following improvements. These would take more time and public review, and are *way* too late for 8.0.

1. By making the **Type** values multi-valued, we preserve more information for users, and for our future use (why the character gets the value it gets). Thus a character could be both "limited-use" (because of the script) and "obsolete" (because in that script it is no longer used).

2. By enhancing the xidmod data to allow multiple characters, we improve the ability to get exactly the desired set of characters. Currently the data file only has information on a per-character basis. That means that it is not possible to indicate that LATIN SMALL LETTER X + COMBINING DOT BELOW is allowed, without also indicating that under-dot is allowed in any combination. This enhancement would allow for the following line in the data file.
   ```
   0078 0323    ; allowed ; recommended # LATIN SMALL LETTER X + COMBINING DOT BELOW
   ```
3. By enhancing the confusable data to allow multiple source characters AND context, we improve the ability to handle confusability for contextual scripts, where two characters might only be confusable in certain contexts. However, this data is trickier to use in implementations, and so we would most likely need to maintain the current file, and have an additional data file with enhanced syntax.