L2/16-335

Title: **A Proposed DerivedName.txt File for the UCD**
Source: Mark Davis (Google, Inc.), Laurențiu Iancu (Microsoft Corporation), and Ken Whistler
Status: Individual contribution
Action: For consideration by the Unicode Technical Committee
Date: 2016-10-29

Proposal

The proposal is to add a new data file DerivedName.txt, as described in this document, to the UCD, in the extracted subdirectory. The file provides a listing of all graphic and format characters with their Name property values.

Background

At the UTC meeting #148, proposal L2/16-187 was discussed and the following action item was assigned:

> B.11.4.1.2 Fixing UAX#44 statement about names property value [Davis, L2/16-187] [148-A13] Action Item for Mark Davis, Laurențiu Iancu: Create a proposed "DerivedName.txt" data file for review at the November 2016 UTC meeting.

This document describes a proposed DerivedName.txt data file and presents a few options for finalizing the structure and contents of the data file.

Proposed DerivedName.txt

The proposed DerivedName.txt file provides a listing of all graphic and format characters assigned in the Unicode codespace, with their Name property values. A prototype DerivedName.txt file, generated for the character repertoire of Unicode 9.0, is attached to this document. The main characteristics of the file are discussed below.

File name and location. The name and location of the file are consistent with those of other derived files in the UCD.

@missing line. The Name property value for unassigned code points (gc=Cn), controls (gc=Cc), private use (gc=Co), surrogates (gc=Cs), and noncharacters (gc=Cn) is <none>. These code points are not listed in the file. An @missing line for the default <none> value is already given in PropertyValueAliases.txt and not duplicated in the proposed DerivedName.txt.

Characters with names generated by rule. There are two options for listing the large sets of characters whose names are generate by rule, such as Han and Tangut ideographs. One option is to list them individually. For example, the characters in the main CJK Unified Ideographs block could be given as:

```
4E00  ; CJK UNIFIED IDEOGRAPH-4E00
4E01  ; CJK UNIFIED IDEOGRAPH-4E01
. . .
9FD5  ; CJK UNIFIED IDEOGRAPH-9FD5
```

The other option for these sets is to list them in a compact form, giving the range of code points and a pattern string for their names. The above set of characters would be given as follows:

```
4E00..9FD5   ; CJK UNIFIED IDEOGRAPH-*
```

Listing each character individually would add about 95,000 lines to the file, for the character repertoire anticipated for Unicode 10.0.

Metacharacter. In the compact format mentioned above, the character '*' is used as a placeholder for the code point. In UAX #42 and the UCD XML data files, the metacharacter used for the same purpose is '#'. However, in the UCD text files, a '#' marks comments, so a different convention is needed to avoid ambiguities.

Hangul syllables. Although the character names of Hangul syllables are defined by rule NR1 of the formal definition of the Name property in the Unicode Standard, they are listed individually in the proposed DerivedName.txt. Only the characters with algorithmic names that include the code point, defined by rule NR2, are listed in compact form.

Other fields. Unlike other derived files in the UCD, which include the General_Category, character counts, and other fields, the proposed DerivedName.txt lists only code points and corresponding Name property values.