L2/14-051

Title: Proposal to coalesce code point ranges in LineBreak.txt and EastAsianWidth.txt
Source: Laurențiu Iancu (Microsoft Corporation) and Ken Whistler (SAP AG)
Status: Individual contribution
Action: For consideration by the Unicode Technical Committee
Date: 2014-01-30

Background

The text files of the Unicode Character Database generally coalesce ranges of characters with the same property values on single lines. Several of the primary UCD files, which are amenable to range grouping given their content, and all of the derived files follow this convention.

A notable exception are the files LineBreak.txt and EastAsianWidth.txt. As of Unicode 6.3, those two files list a single character per line, in the vast majority of cases, although there are long ranges of consecutive code points which are assigned the same Line_Break or East_Asian_Width property values, respectively. Additionally, the two files do include a few ranges, but those use a legacy format, which is only used in UnicodeData.txt elsewhere in the UCD and documented, specifically for UnicodeData.txt, in Section 4.2.3, *Code Point Ranges* of UAX #44.

During the property definition work for Unicode 7.0, the initial drafts of LineBreak.txt and EastAsian-Width.txt each have received an additional 2,833 lines for as many newly-encoded characters and grew to over 27,000 lines. Both files contain long sequences of lines which can be efficiently coalesced to single lines using code point ranges. For example, a section from LineBreak.txt such as

```
A016;ID # YI SYLLABLE BIT
A017;ID # YI SYLLABLE BIX
...
A48C;ID # YI SYLLABLE YYR
```

can be equivalently formatted as a single line as

```
A016..A48C    ; ID # Lo [1143] YI SYLLABLE BIT..YI SYLLABLE YYR
```

the same way as, for instance, in DerivedLineBreak.txt. Similar reformatting would change the legacy-style code point range lines from the "<*schematic name*, First>..<*schematic name*, Last>" pattern

```
3400..4DB5;ID # <CJK Ideograph Extension A, First>..<CJK Ideograph Extension A, Last>
```

to the more common "*character name*..*character name*" pattern

```
3400..4DB5    ; ID # Lo [6582] CJK UNIFIED IDEOGRAPH-3400..CJK UNIFIED IDEOGRAPH-4DB5
```

Such changes would both compress the two files and format them more consistently with other UCD files, as well as bring them in alignment with the documentation in UAX #44 without further editing. Parsers of LineBreak.txt and EastAsianWidth.txt should be able to handle the condensed format because ranges are already present in both files, besides being used in other UCD files.

Proposal

To reduce the size of the UCD files LineBreak.txt and EastAsianWidth.txt and format them similarly to other UCD files, the proposal is to apply the following changes for Unicode 7.0:

1. Replace the contiguous ranges of individually listed code points which are assigned the same property values with single lines, following the format of other UCD files, as shown earlier, and
2. Reformat the few existing ranges which use the legacy "<*schematic name*, First>..<*schematic name*, Last>" pattern to the more common "*character name*..*character name*" pattern used in the other UCD files except UnicodeData.txt.