L2/19-374

Re: Data file format issues

From: Mark Davis

To: UTC

Date: 2019-10-29

---

1. Several people at the recent conference discussed a technique that they had for dealing with the fact that data files in the UCD may change location. What they do is to ignore the file structure, simply recursively search for a file wherever it might be. So that this technique continues to work in the future (we are moving files in that way in Unicode 13.0, for example), I propose that we minute and document the following:
   - No Unicode data files, in the UCD or associated with other technical standards, will be given the same name, except for documentation files in data directories, such as a ReadMe.txt.
   - CLDR is an exception: for example, multiple files with the <ldml> DTD are logically considered to be a whole. For example, annotations/en.xml, annotationsDerived/en.xml, collation/en.xml, … are all logically part of the <ldml> file for English.
2. The format of multivalued files makes it really hard to diff files see what values change over versions. I'd recommend that for Identifier_Type, Script_Extensions, and any other multivalued properties like Script Exceptions, we sort by code point without grouping by the property values. That should have no effect on any parsers, but make it diffable. Since that has no semantic difference, we could apply it for this release.

That is, instead of

```
…
#  IdentifierType:  Exclusion Obsolete Not_XID

2CF9..2CFC    ; Exclusion Obsolete Not_XID # 4.1    [4] COPTIC OLD NUBIAN FULL STOP..COPTIC OLD NUBIAN VERSE DIVIDER
…
```

We'd see a list in code point order, with one or more values on each line, as follows.

(Right now, the values are sorted by "importance". We could sort alphabetically instead, for better predictability.)

```
0009..000D    ; Not_XID                                 # 1.1    [5] <control-0009>..<control-000D>
0020..0026    ; Not_XID                                 # 1.1    [7] SPACE..AMPERSAND
```

```
0027          ; Inclusion                          # 1.1        APOSTROPHE
0028..002C    ; Not_XID                            # 1.1    [5] LEFT PARENTHESIS..COMMA
002D..002E    ; Inclusion                          # 1.1    [2] HYPHEN-MINUS..FULL STOP
002F          ; Not_XID                            # 1.1        SOLIDUS
0030..0039    ; Recommended                        # 1.1   [10] DIGIT ZERO..DIGIT NINE
003A          ; Inclusion                          # 1.1        COLON
003B..0040    ; Not_XID                            # 1.1    [6] SEMICOLON..COMMERCIAL AT
0041..005A    ; Recommended                        # 1.1   [26] LATIN CAPITAL LETTER A..LATIN CAPITAL LETTER Z
005B..005E    ; Not_XID                            # 1.1    [4] LEFT SQUARE BRACKET..CIRCUMFLEX ACCENT
005F          ; Recommended                        # 1.1        LOW LINE
0060          ; Not_XID                            # 1.1        GRAVE ACCENT
0061..007A    ; Recommended                        # 1.1   [26] LATIN SMALL LETTER A..LATIN SMALL LETTER Z
007B..007E    ; Not_XID                            # 1.1    [4] LEFT CURLY BRACKET..TILDE
0085          ; Not_XID                            # 1.1        <control-0085>
00A0          ; Not_XID Not_NFKC                   # 1.1        NO-BREAK SPACE
00A1..00A7    ; Not_XID                            # 1.1    [7] INVERTED EXCLAMATION MARK..SECTION SIGN
00A8          ; Not_XID Not_NFKC                   # 1.1        DIAERESIS
00A9          ; Not_XID                            # 1.1        COPYRIGHT SIGN
00AA          ; Not_NFKC                           # 1.1        FEMININE ORDINAL INDICATOR
00AB..00AC    ; Not_XID                            # 1.1    [2] LEFT-POINTING DOUBLE ANGLE QUOTATION MARK..NOT SIGN
00AD          ; Not_XID Default_Ignorable          # 1.1        SOFT HYPHEN
00AE          ; Not_XID                            # 1.1        REGISTERED SIGN
00AF          ; Not_XID Not_NFKC                   # 1.1        MACRON
00B0..00B1    ; Not_XID                            # 1.1    [2] DEGREE SIGN..PLUS-MINUS SIGN
00B2..00B4    ; Not_XID Not_NFKC                   # 1.1    [3] SUPERSCRIPT TWO..ACUTE ACCENT
...
FE27..FE2D    ; Technical                          # 7.0    [7] COMBINING LIGATURE LEFT HALF BELOW..COMBINING CONJOINING MACRON BELOW
FE2E..FE2F    ; Uncommon_Use Technical             # 8.0    [2] COMBINING CYRILLIC TITLO LEFT HALF..COMBINING CYRILLIC TITLO RIGHT HALF
FE30..FE32    ; Technical Not_XID Not_NFKC         # 1.1    [3] PRESENTATION FORM FOR VERTICAL TWO DOT LEADER..PRESENTATION FORM FOR VERTICAL EN DASH
FE33..FE34    ; Technical Not_NFKC                 # 1.1    [2] PRESENTATION FORM FOR VERTICAL LOW LINE..PRESENTATION FORM FOR VERTICAL WAVY LOW LINE
FE35..FE44    ; Technical Not_XID Not_NFKC         # 1.1   [16] PRESENTATION FORM FOR VERTICAL LEFT PARENTHESIS..PRESENTATION FORM FOR VERTICAL RIGHT WHITE
    CORNER BRACKET
FE45..FE46    ; Technical Not_XID                  # 3.2    [2] SESAME DOT..WHITE SESAME DOT
...
```