L2/11-183

To: UTC
From: CLDR-TC
Date: May 10, 2011
Subject: Embedding Level Mark

Draft (http://goo.gl/xNFJo)

In CLDR we've realized that there are a number of instances where people want a format to be composed of fields that flow in the paragraph direction. For example, people want date formats that order one direction in a RTL environment, but in the opposite direction in a LTR environment.

To solve that need, the CLDR-TC committee is recommending that the UTC add a new BIDI ordering mark

U+XXXX Embedding Level Mark (ELM)

The semantics are the same as the LRM and RLM except that the character behaves as if it were an LRM whenever the Bidi embedding level is L, and behaves as if it were an RLM whenever the Bidi embedding level is R. A possible code point for XXX would be 2065, which is a **Default_Ignorable_Code_Point.**

This can then be used as follows. Suppose, for example, that you have a neutral surrounded by numbers: 12/34

This appears as 12/34 in either RTL or LTR environments. By inserting an ELM before the '/', it would appear as 12/34 in a LTR environment, and as 34/12 in a RTL environment.

The new character would have bidi class ON. There would be a new rule:

W0. Examine each embedding-level character (ELM) in the level run, and set the bidi type to L if the level is even, and R if the level is odd.

Notes

1. It would have been preferable to define a new bidi class for this ELM behavior; it could then be used as a bidi class override, which—in situations that permitted such overrides—could achieve the ELM behavior without insertion of extra mark characters. However, per the Unicode Character Encoding Stability Policy, "The **Bidi_Class** property values will not be further subdivided." [http://www.unicode.org/policies/stability_policy.html#Property_Value]

2. With this mark, as with LRM and RLM, it would be useful to provide a recommendation for placement: When the mark is intended to modify the class of an otherwise-neutral character, should the mark be placed before or after the character? The mark placement affects how the neutral character interacts with other nearby characters per UAX#9 rules W5 and W4.

Relationship to other proposals

1. **L2/11-005,** “Proposal to encode an Arabic-Letter Mark (ALM),” Matitiahu Allouche, Mohamed Mohie: The proposed ALM is like RLM but with class AL instead of R; and thus with a different effect on the behavior of succeeding digits. It addresses a different set of problems than the ELM, and seems quite useful in its own right as a separate character; there is no need to unify that proposal with this one.

2. **L2/10-200,** “Tailoring the Unicode Bidi Algorithm,” Murray Sargent: This is a nice overview of several different issues. The section “Internationalized Resource Identifiers” describes issues and solutions that overlap with those addressed by this ELM proposal. Murray notes that a strategy adopted by RichEdit for display of identified IRIs is to “force the delimiters '#', '.', '/', ':', '?', '@', '[', ']' to follow the paragraph (or embedding) direction. … This approach appears to be ideal. The only problem is that it’s not trivial to identify IRIs using heuristics.”

Using the ELM adjacent to the separators might be a way to achieve the same result in a way that is portable across applications, not all of which will necessarily perform the IRI identification and direction-class forcing described in L2/10-200.

Feedback (an earlier version of this was sent to the bidi list on Apr 28):

1. From Mati Allouche (on bidi list):

> I understand the need for such behavior. I don't like it being implemented by a character with a special rule. As far as I know, that would be a first for the UBA.
>
> The UBA is expressed in terms of bidi classes, not of characters. Even LRE, RLE, LRO, RLO, PDF are treated as bidi classes, although they only include one character per class and have very little chances to ever need to add other characters in these classes.
>
> Therefore I suggest to create a new bidi class called for instance Embedding Level Class (ELC) and have the new character ELM have the bidi class ELC.
>
> One application of the ELC could be as follows. Instead of inserting an ELM in the middle of 12/34, or around separators in IRIs, we could override the bidi class of separators (the slash for 12/34; the dot, commercial at etc... for IRIs) when applying the UBA to those special strings and avoid adding ELMs in the middle of the data.
>
> Note that the bidi engine in ICU4C and ICU4J allows the user to override the bidi class of any character.

2. From Murray Sargent (response to direct e-mail asking for comments):

> (paraphrasing) Murray indicated that he thought the ELM was a good idea, and would be easy to implement.