Title: **Regularize Extended_Pictographic**
From: **Mark Davis, CLDR-TC**
To: **UTC**
Date: **2017-03-22**
Draft: **link**

We created extended segmentation rules in CLDR to handle emoji when we didn't have time to make the right changes in UAX#14 and UAX#29. That means that anyone who wants to properly handle Emoji in segmentation must use the CLDR additions: customized LDML rules and the additional CLDR data in Extended_Pictographic.txt. Since we'd like everyone to be future-proofed, that effectively means that everyone should ignore the plain rules and use the CLDR ones (which also go into ICU).

Once E5.0 (UTS #51 Emoji 5.0) is approved, we can fix that situation in subsequent versions of the Unicode Standard and CLDR. This is probably too substantial for people to handle in U10.0, but we can have a clear direction for the future. *However, it would be worth putting out a PRI early to get feedback on this.*

Proposal

UTS #51

Move **Extended_Pictographic** from CLDR into the Emoji data files, for the next version of UTS #51 after Emoji 5.0 (Emoji 6.0 or perhaps a sooner small update Emoji 5.1, whatever timing is needed).

That is, add data for new Emoji property with the following contents:

> Extended_Pictographic =
> Extended_Pictographic.txt (from CLDR)
> + [:Emoji:]
> - [:Emoji_Component:]

- Note that for each new version of E5.0, we'd need to review any recently-added Unicode symbols to see whether they should also be added.

CLDR

Deprecate **Extended_Pictographic.txt** and remove the data (leave a readme in the data file).

In LDML, point to the emoji data for the replacement values for Extended_Pictographic.txt, and remove the customized rules that override UAX#14 and UAX#29.

UAX#14 and UAX#29

Add text along the lines of the following:

> The rules for segmentation may may be augmented by use of UnicodeSet notation using properties or literals outside of the main property associated with the algorithm, such as [:General_Category=Letter:] or [\uFE0F]. The properties may either be UCD character properties, or properties associated with a UTS such as *UTS #51, Unicode Emoji*.

Modify the segmentation rules in UAX#14 and UAX#29 based on LDML (updated somewhat). The *old rules* are presented for comparison, while the new rules are presented with a prime′ mark.

Grapheme Cluster Break

GB10	*(E_Base | EBG) Extend**	×	*E_Modifier*

GB10′	[:Emoji_Modifier_Base:] [**\uFE0F]***	×	[:Emoji_Modifier:]
GB11	*ZWJ*	×	*(Glue_After_Zwj | EBG)*
GB11′	[:Extended_Pictographic:] [**\uFE0F]?** ZWJ	×	[:Extended_Pictographic:]

Word Break

WB3c	*ZWJ*	×	*(Glue_After_Zwj | EBG)*
WB3c′	ZWJ	×	[:Extended_Pictographic:]
WB14	*(E_Base | EBG)*	×	*E_Modifier*
WB14′	[:Emoji_Modifier_Base:]	×	[:Emoji_Modifier:]

Line Break

LB8	*ZWJ*	×	*(ID | EB | EM)*
LB8a′	ZWJ	×	(ID | [:Extended_Pictographic:])
LB30b	*EB*	×	*EM*
LB30b′	[:Emoji_Modifier_Base:]	×	[:Emoji_Modifier:]

Outside of the listed rules LB8 and LB30b, the property values EB and EM behave like ideographic characters, LB class ID. This shows up in the following.

LB22	(ID | EB | EM)	×	IN
LB23a	PR	×	(ID | EB | EM)
	(ID | EB | EM)	×	PO

UCD

The characters with the property values in UAX#14 and UAX#29 that were added for emoji — except for ZWJ — can be either left alone (their properties wouldn't be necessary any more for preventing emoji breaks), or changed to Other (for GCB / WB) or Ideographic (LB).

Leaving them alone would be the minimal change, while changing their values would allow the rules to be simplified by changing (ID | EB | EM) → ID in LB 22 and LB 23a.

Here are the changed values :

Property	U10.0 Value	U8.0 Value	Characters
Grapheme_Cluster_Break	E_Base	Other	[-□-□]
Grapheme_Cluster_Break	E_Base_GAZ	Other	[-]
Grapheme_Cluster_Break	E_Modifier	Other	[-]
Grapheme_Cluster_Break	Glue_After_Zwj	Other	[♀ ♂]
Word_Break	E_Base	Other	[-□-□]
Word_Break	E_Base_GAZ	Other	[-]
Word_Break	E_Modifier	Other	[-]
Word_Break	Glue_After_Zwj	Other	[♀ ♂]
Line_Break	E_Base	Ideographic	[ ]
Line_Break	E_Base	Unknown	[-□-□]
Line_Break	E_Modifier	Alphabetic	[-]