# Unicode Support

Michael Ficarra • 76th Meeting of TC39, June 2020

# History

## July, 2016

Committee approves normatively referencing "latest" Unicode version.

### 10.i.a Require Unicode 9.0.0

Brian Terlson

Patch: Require Unicode 9.0.0 instead of 8.0.0

BT: We can update every year, or change language

EFT: "use latest"

AWB: Normatively referencing a spec with no version means "the latest"

BT: There are no changes in 9.0.0. Committed to reviewing changes each year.

RW: The 402 editor will have to do the same task

(agreement)

WH: New Unicode case mappings might cause \p{} (and especially ^\P{}) problems analogous to the \W problem we recently encountered.

BT: This won't be added in the future, but they can't be "fixed" because they can't break compatibility.

DT: Should the requirements be clear for ES implementations?

BT: risk with respect to breaking change updating Unicode, larger possibility with \p

- Unicode is reluctant to touch case mapping
- · characters might move

DE: But these tend to be bug fixes?

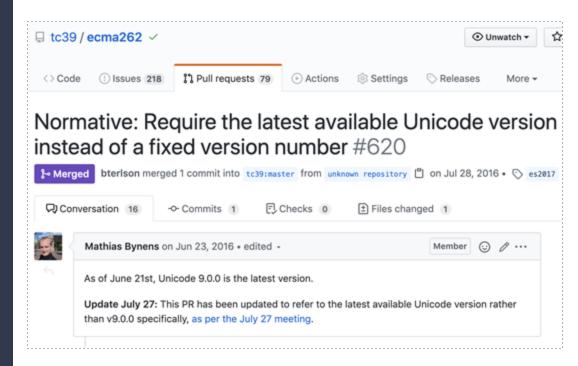
BT: Still a risk of a breaking change

#### Conclusion/Resolution

- · Do not merge the PR
- Use language for "the latest", i.e. unbounded reference
- · This should be applied to 402 as well

# July, 2016

Editor merges PR to change wording

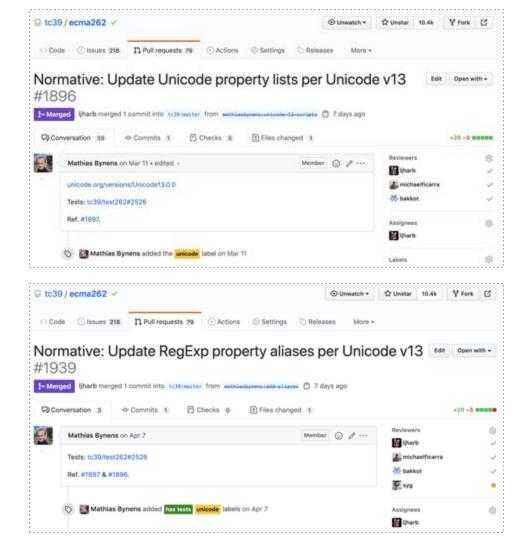


# Done forever, right?

### 2020

Still making regular Unicode version bump PRs.

What gives?



## Tables 55, 56, 57

Used to dictate what's allowed in RegExp  $\$ \p{...}/.

Property names, values, and their aliases.

Table 55: Non-binary Unicode property aliases and their canonical property names

Property name and aliases	Canonical property name
General_Category	General_Category
gc	
Script	Script
sc	
Script_Extensions	Script_Extensions
scx	

#### Table 56: Binary Unicode property aliases and their canonical property names

Property name and aliases	Canonical property name
ASCII	ASCII
ASCII_Hex_Digit	ASCII_Hex_Digit
AHex	
Alphabetic	Alphabetic
Alpha	
Any	Any
Assigned	Assigned

### Table 57: Value aliases and canonical values for the Unicode property General\_Category

Property value and aliases	Canonical property value
Cased_Letter	Cased_Letter
LC	
Close_Punctuation	Close_Punctuation
Pe	
Connector_Punctuation	Connector_Punctuation
Pc	
Control	Control
Cc	

## Constraints

- Don't want to pull in all Unicode properties.
- Don't want observably mixed Unicode support required for conformance.
- Don't want to maintain big tables of Unicode properties, if we can avoid it.

## Proposal

- Normatively describe the process for deriving the information currently in tables 55, 56, and 57 from a Unicode data set.
- Commit to maintaining a consistent observable Unicode version.
- Optionally, provide non-normative tables that will become inaccurate following a Unicode release. This might be helpful for implementers?
- Do this as a normative PR.