

Unicode Support

Michael Ficarra • 76th Meeting of TC39, June 2020

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

History



July, 2016

Committee approves normatively referencing "latest" Unicode version.

10.i.a Require Unicode 9.0.0

Brian Terison

[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

The screenshot shows a GitHub pull request interface for the repository `tc39 / ecma262`. The title of the pull request is "Normative: Require the latest available Unicode version instead of a fixed version number #620". It indicates that the pull request has been merged. Below the title, there is a section for conversation, showing a comment by Mathias Bynens from June 23, 2016, which was edited. The comment states that as of June 21st, Unicode 9.0.0 is the latest version. It also includes an update from July 27, stating that the pull request has been updated to refer to the latest available Unicode version rather than v9.0.0 specifically, as per the July 27 meeting. The interface includes navigation tabs for Code, Issues (218), Pull requests (79), Actions, Settings, Releases, and More. It also shows statistics for Conversation (16), Commits (1), Checks (0), and Files changed (1).

tc39 / ecma262 ✓


Unwatch

<> Code ⓘ Issues 218 🏹 Pull requests 79 ⚙️ Actions ⚙️ Settings ⌚ Releases More ▾

Normative: Require the latest available Unicode version instead of a fixed version number #620

🔗 Merged bterlson merged 1 commit into `tc39:master` from `unknown repository` on Jul 28, 2016 • es2017

💬 Conversation 16 🔗 Commits 1 📋 Checks 0 📄 Files changed 1

 **Mathias Bynens** on Jun 23, 2016 • edited ▾ Member 😊 ✎ ⋮

As of June 21st, Unicode 9.0.0 is the latest version.

Update July 27: This PR has been updated to refer to the latest available Unicode version rather than v9.0.0 specifically, [as per the July 27 meeting](#).

Done forever, right?



2020

Still making regular Unicode version bump PRs.

What gives?

This screenshot shows a GitHub pull request titled "Normative: Update Unicode property lists per Unicode v13" with the number #1896. The repository is tc39/ecma262. The pull request is merged and was created 7 days ago. The commit message is "unicode.org/versions/Unicode13.0.0". The tests are "tc39/test262#2526" and the reference is "#1897". The pull request was reviewed by @harb, @michaelficarra, and @bakkot, all of whom gave their approval. The pull request was assigned to @harb. The pull request was labeled with the "unicode" label.

This screenshot shows a GitHub pull request titled "Normative: Update RegExp property aliases per Unicode v13" with the number #1939. The repository is tc39/ecma262. The pull request is merged and was created 7 days ago. The commit message is "mathiasbynens-add-aliases". The tests are "tc39/test262#2526" and the references are "#1897 & #1896". The pull request was reviewed by @harb, @michaelficarra, @bakkot, and @syg, all of whom gave their approval. The pull request was assigned to @harb. The pull request was labeled with the "has tests" and "unicode" labels.

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	
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.