Ecosystem Infra 2017 Q4 OKRs

Overall Q4 score: 0.36 priority weighted

The team's mission is to *establish the infrastructure and culture that empower web platform engineers to make platform-wide changes with low friction.*

We prioritize work according to the expected benefit to the ecosystem. Rules of thumb:

- Risk is OK. A 50% chance of impact 2*x* is just as good as certainty of impact *x*.
- Long term is OK, projects can have a high expected impact without being urgent.
- Because of the day-to-day value and compounding effects, launching early and maintaining momentum matters.

Priority and scoring:

- The priority levels are P1, P2 and P3. (P0 is not used.)
- The priority of KRs are "global," i.e., all P1 KRs are of higher priority than all P2 KRs.
- The scoring should be such that averaging 0.7 is expected, and 1.0 is exceptional.

Mid-Q check-in was on 2017-11-14, i.e. data was up to and including 2017-11-13, 45 days.
Green means on track and likely to score ≥0.7, Yellow means on track but with some risk, Red means at risk and unlikely to score ≥0.7.

The OKRs

O: web-platform-tests are a first-class citizen in Blink (P1)

Owner: robertma

KR: WPT import latency is less than 12 hours (P1)

Owner: robertma, rakuco
Import latency is the time between the original PR is merged on GitHub and the import CL that contains the PR lands in Chromium. Measured by wpt-import-stats.py.
Scoring: percentage of imports < 12hrs
Reference: Q3 P1 KR: automatic import happens once a day (~= 24hr latency). Sept (as of 25th) ~56%. There is reasonable hope to reach 60% to 70% in Q4.
Mid-Q: Red: 0.31 See this thread for more details on the breakage in Oct
End-Q: 0.43 (We got 84% in the smooth December!)

KR: WPT export latency is less than 1 hour (P1)

Owner: robertma

Export latency is the time between the original CL lands in Chromium and the corresponding export PR is merged on GitHub. Measured by wpt-export-stats.py.

Scoring: let p = percentage of exports < 1hr, score = (p-0.5)*2

Reference: Q3 SLA: all exports within 24hrs and 50% exports within 25mins. Aug & Sept (as of 25th), ~85% (score = 0.7), which should be at least kept in Q4.

Mid-Q: Green: 0.66 (83%)

End-Q: 0.75 (87%)

KR: Consolidate CSS test suites into a single unversioned directory per spec (P1)

Owner: gsnedders

Rationale: This is important for the style and layout teams, as the versioned directories make it harder to work with wpt than LayoutTests. Needs to be fast to unblock their further investigation.

Scoring: Based on when #7503 is fixed. 0.7 if done on October 20, 0.0 by end of Q4. (Extrapolate to >0.7 if done early.)

Mid-Q: Done: 0.59. (Issue is still open, but time of big rename is what counts)

KR: WPT import notifications (P2)

Owner: robertma

Scoring:

- +0.3: Wrote a design doc and reached some consensus among stakeholders (test owners)
- +0.2: Implemented the system
- +0.1 for each top-level directory opting in (candidates: IndexedDB, css, html, dom, fetch, streams), ceiled at 1.0

Mid-Q: Red: 0.0. Not started, but still in the plan.

2017-11-22: design doc sent

End-Q: 0.8 (0.3+0.2+0.1*3: infrastructure, fullscreen, flexbox)

Also, here's a tracking spreadsheet for the enrolment

KR: Resolve long-standing issues that affect the reliability of WPT in Blink (P2)

Owner: robertma, rakuco

Scoring: +0.2 for each of the item in the following list clamped to 1.0

- Fixed screenshot timing issue in ref/pixel tests with web fonts (https://crbug.com/507054)
- Investigated the use of Apache in LayoutTests and the viability to get rid it
- Investigated supporting LayoutTests (including WPT) in Chromium FindIt, and the possibility to use it for better detection of flakiness
- Document importer & exporter (https://crbug.com/756216)
- Wrote an audit tool for Chromium exports (https://crbug.com/754619)
- Simplify Ecosystem Infra Rotation playbook (make manual imports easier, automate checking the health of importer/exporter, etc.)

Mid-Q: Yellow: 0.4 (FindIt, import/export docs almost done)

End-Q: 0.6 (underlined items done)

KR: 50% of Blink changes using layout tests are using web-platform-tests (P3)

Owner: robertma, rakuco
Scoring: let p = actual percentage, t = target percentage, score = p/t
Reference: The number has been around 20% throughout the year, with September seeing a positive trajectory (23%, but it's also worth noting that the total number of source+test changes is much smaller this month, so the stat might be less representative).
Note: robertma will maintain the script, keep track of the data and report when something interesting pops out in the data. Some outreach will be needed to actually push the number.
Mid-Q: Red: 0.33. 16% of source+test changes in Q4 so far use WPT.
End-Q: 0.38 (19%)

O: web-platform-tests are a first-class citizen in the standards process (P2)

KR: A policy for web-platform-tests coverage is adopted for 40 more specs (P2)

Owner: foolip, rakuco
Scoring: Linearly based on progress towards 40.
Progress: 2017 Q4 testing policy progress
Mid-Q: Green: 39/40 = 0.975
End-Q: 39/40 = 0.975

KR: IDL from specs is automatically imported into web-platform-tests (P3)

Owner: markdittmer, gsnedders
Scoring:

- 128 is the number of current tests that use idlharness.js on 2017-10-11
- N = number of tests converted to be auto-updated (even if no spec change)
- Score is N/128

Mid-Q: Red: No progress on this; time has been spent on P<=2. Cut for Q4.
End-Q: 0.0

Standalone IDL files are in /interfaces/*.idl. More tests to have to converted to the standalone style to meet the KR.

KR: Determine test coverage for key web platform specs (P2)

Owner: tabatkins, gsnedders
Rationale: Relates to the “Support linking to tests from specs” KR. In addition to doing the linking, also write up what work needs to be done for key specs to get them close to 100% test coverage. Candidate specs are CSS Flexbox, Grid, Align and Sizing + DOM.
Scoring: +0.2 for each spec for which we have such a report.
Mid-Q: Red: Depends on the Bikeshed feature, which is on good track, but it’s a stretch that 4 (or 5) specs would be fully annotated such that we know the coverage in detail.

End-Q: 0.0

O: Confluence is an attractive tool for prioritizing work on all browsers (P2)

Owner: markdittmer

KR: Close all urgent Confluence bugs filed before Nov 1 (P1)

Rationale: Data issues that matter to TLs that are actively using the tool will be addressed within the first month. Urgent issues.
Scoring:

- Measure on Nov 1: A = Issues(label:urgent)
- Measure at EoQ: B = A(is:closed)
- Score: Count(B) / Count(A)

Mid-Q: Green: Nearly all closed. Remaining 1 is partially resolved, but require more legwork to reflect browsers' reality more accurately.
End-Q: 0.78 (2 / 9 issues left open)

KR: New browser releases are available within 7 days (P2)

Scoring: Take average of time-to-first-data-push after a release of each browser (Chrome 62, Edge 16, Firefox 57, Safari 11) becomes available on BrowserStack this quarter:

- +0.5 < 14 days
- +0.1 < 12 days
- +0.1 < 10 days
- +0.1 < 9 days
- +0.1 < 8 days
- +0.1 < 7 days

Mid-Q: Yellow: Made progress on release process itself, but BrowserStack often lags substantially behind official release dates.
End-Q: 0.5 maybe?

KR: Confluence TTI < 3s (P2)

Scoring: Average computed from view rendered when clicking the last point on all confluence graphs.

- +0.5 for monitoring the TTI metric at all
- +0.1 Average(TTI) < 12s
- +0.1 Average(TTI) < 10s
- +0.1 Average(TTI) < 7s
- +0.1 Average(TTI) < 5s
- +0.1 Average(TTI) < 3s

Mid-Q: Green/Yellow: Seems on good track, need to decide whether we want to push for full performance win or just stay where we are.
End-Q: 0.6 (Empty cache: 14606ms, Full cache: 5990ms; Avg of two: 10298.5ms < 12s)

O: Increase the quality of web-platform-tests submissions (P2)

Owner: lukebjerring

KR: Surface actionable test results as GitHub review comments (P1)

Owner: lukebjerring
Scoring: should fix "Surface test regressions from Travis as review comments"

- +0.5 for surfacing diffing endpoint
- +0.5 for adding comments for regressions

Mid-Q: Yellow: Making progress on diffing calculation but not comments
End-Q: 0.5

KR: Surface actionable test results as Gerrit review comments (P3)

Owner: lukebjerring
Mid-Q: Red: No work done yet, nothing to do until the above KR is complete.
End-Q: 0

O: Support major web-platform-tests automation use cases (P1)

Owner: kereliuk, gsnedders

KR: Land click test automation using WebDriver in wpt (P1)

Owner: gsnedders
Rationale: Carried over from "Infrastructure for writing WebDriver tests landed in wpt" and "Infrastructure for passing user gesture requirements landed in wpt" in KRs in Q3, both of which will be resolved by landing https://github.com/w3c/web-platform-tests/pull/6897.
Scoring: Based on time of landing. Linear between 1.0 on Oct 5 and 0.0 on Oct 20.
Mid-Q: Red: 0.0, done 2 Nov

> Philip: In hindsight it was a mistake to let the score go to 0 before the end of the quarter, as it makes the mid-quarter check-in pointless. This was urgent so some sort of deadline was appropriate, but a score of 0.5 would have been more fair since it was done.

KR: Land keyboard input automation using WebDriver in wpt (P2)

Owner: gsnedders, kereliuk
Rationale: Required for Upstream focus navigation related layout tests to WPT
Scoring: Based on time of landing. Linear between 1.0 on Oct 23 and 0.0 at Dec 22.
Mid-Q: Yellow: https://github.com/w3c/web-platform-tests/pull/8159
End-Q: 0.0, Decided to abandon the above PR to use the webdriver API directly instead of selenium

KR: Bypass permissions and use mock media for WebRTC on wpt.fyi (P1)

Owner: kereliuk
Rationale: See GitHub: Pass browser-specific command line flags when running webrtc/
Scoring: Based on number of browsers passing a simple test involving getUserMedia() in http://wpt.fyi/webrtc, if that browser+version has command line flags to do it.
(Not blocked on WebRTC WebDriver extension in spec.)

- +0.25 for each browser done (Edge very unlikely)
- +0.25 if any browser done before TPAC

Mid-Q: Red: 0.0 spent a lot of time on this and have not made much progress. Can someone help me with this or should this be reprioritized?
End-Q: 0.0 Spend a lot of the quarter working on this and ran into a lot of bugs with the dashboard and had a lot of issues reproducing behaviour. Did not get any browser working on wpt.fyi.

KR: Implement the Permissions API WebDriver extension in ChromeDriver (P2)

Owner: kereliuk
Scoring:

- +0.5 if CL exists
- +0.2 if landed
- +0.3 if some test is converted and passing on wpt.fyi

Mid-Q: Yellow: 0.0

DevTools also wants this regardless of WebDriver, going to implement. @lushnikov has this on his queue for DevTools, I've told him to add me as a reviewer and to come to me if he needs help or me to take it over. Will follow up on Friday. Permissions PR was pushed to TPAC, WG (including me) says they will review extension spec

End-Q: 0.0, Not completed

KR: Spec for mock media and implementation design doc (P3)

Owner: kereliuk
Scoring:

- +0.2 if spec exists
- +0.5 for design doc
- +0.3 for implementing (unlikely, stretch)

Mid-Q: Red: 0.0 Haven't started
End-Q: 0.0, Not completed

KR: Almost all new tests are automated in (upstream) WPT (P3)

Owner: kereliuk
Scoring:

- Let N be the number of new tests written between Nov 1st and Dec 31
- Let M be the number of manual tests written between Nov 1st and Dec 31

- Score = max{0, (1 - M/N * 10)}

Mid-Q: If we go from October 3rd

- November 27th
  - 53389 tests
  - 3237 manual tests
- October 3rd
  - 60654 tests
  - 3222 manual tests

So it seems the total number of tests went down, I'm not sure of an easy way to count the number of new tests written after a certain day. We can't get this from the manifest file, so we would need to look at all commits and do something smart with that.
End-Q: not scored

KR: Sort existing manual tests and plan for automating them/outsourcing (P3)

Owner: kereliuk, gsnedders
Scoring:

Old Scoring:

- Start at score 1
- Let T be a date
- -0.01 for every manual test without a plan for automation at the end of T

New Proposed Scoring

- +0.3 mapping every untestable flag in wpt to a manual test which can be automated given that capability claiming none exist
- +0.7 for a mapping of every manual test to a bug tracking either its untestability or an action to automate it (scored as a percentage)

Mid-Q: Red: 0.0 have not automated any manual tests yet, but collecting data and starting to build automating off gsnedders testdriver.js
End-Q: 0.96 Tracked here. No mapping for 63 of 1132 manual tests (omitted svg/import and css/)

Info

O: ChromeDriver is a well-engineered piece of software (P2)

KR: ChromeDriver has a canary/nightly release channel (P1)

Owner: kereliuk, zhanliang
Scoring:

- +0.4 for a CL
- +0.3 for landing the CL

- +0.3 for modifying wpt to use nightly in CI

Mid-Q: Yellow: 0.3, a CL exists here for making the download the same process and location and official releases. All CI builds are completely available actually and linked on the website, but it is cumbersome to download
Mid-Q: Green 0.7, Is public and linked on the ChromeDriver website, but WPT is not modified for its use

The reason I have chosen only canary/stable not dev and beta is for simplicity and impact. There are developers who use canary Chrome, and hence also should be using the cutting edge of ChomeDriver since it often contains bug fixes introduced by the canary binary. For example: https://github.com/w3c/web-platform-tests/issues/6986
For simplicity since we don't sync with any of the chrome release channels currently, I think it makes most sense to do just stable and canary to start, and later we can evaluate if it is useful to do beta and dev.

KR: ChromeDriver is highly interoperable with other WebDriver implementations (P3)

Owner: kereliuk, zhanliang
Scoring: Score is percentage of pass rates from WPT webdriver suite
Mid-Q: Yellow: 0.3 wpt.fyi currently not working and webdriver tests start hanging randomly (will look into this), but last I remember was around 30%

KR: W3C flag is replaced with a "legacy" flag for old behavior that is not default (P2)

Owner: kereliuk, zhanliang
Scoring: +0.2 for a CL with a warning message of a switch date, +0.5 for CL, +0.3 for version release
Mid-Q: Yellow: I'm predicting I can finish this before end-Q, but right now score is 0.0.
End-Q: 0.0. Did not do

KR: Understand the blockers for two-way communication WebDriver and ChromeDriver (P3)

Owner: kereliuk
Scoring:

- +0.5 a doc discussing potential implementations
- +0.2 for discussing how an implementation would be possible in ChromeDriver
- +0.3 for discussing how this could be standardized

Mid-Q: Yellow: 0.2 From the discussions I've had with the WG
End-Q: 0.2, no further work done after TPAC

KR: Simple daily usage counting is collected (P2)

Owner: kereliuk
Scoring:

- +0.2 for a design doc
- +0.1 for a privacy review
- +0.2 for implementation
- +0.5 for a working implementation in a chromedriver release

Mid-Q: Yellow: 0.2 Did preliminary reading, need to get back to this
End-Q: 0.0. Did not do

KR: Endpoint metrics are collected (P3)

Owner: kereliuk
Scoring:

- +0.2 for a design doc
- +0.1 for a privacy review
- +0.2 for implementation
- +0.5 for a working implementation in a chromedriver release

Mid-Q: Yellow: 0.2 Did preliminary reading, need to get back to this
End-Q: 0.0. Did not do

O: Upstream web-platform-tests infrastructure is reliable and maintainable (P2)

KR: web-platform-tests pull requests have test results from Chrome, Edge, Firefox and Safari within 15 minutes in Nov-Dec (P1)

Owner: mattl
Rationale: This is a proxy for the reliability of the TravisCI infrastructure. Errors in running tests will show up as a reduction in this metric.
Scoring: For each PR that should run tests, determine the test result latency for each browser. Score as 1 if the latency is < 15 minutes and otherwise 0. The total score is the average of these per-PR-and-browser scores. This is calculated at https://pulls.web-platform-tests.org/performance?start=2017-11-01&end=2017-12-31
Application logs are also available on this server at /var/www/wptdash/site/logs/access.log, which you can grep for transactions to /api/pull and /api/build to identify the time of application interactions between pull requests and travis builds.
Mid-Q: Red: TODO(mattl): write scripts to calculate and add to foolip/ecosystem-infra-stats
End-Q: 0.53

KR: Include results from recent Edge and Safari versions (P2)

Owner: mattl
Scoring:

- +0.4 for each of Edge 16 and Safari 11
- +0.1 for each of Edge Windows Insider and Safari Technology Preview

Mid-Q: Red: recent PR using Ege 14 and Safari 10. Some work to be done regarding BrowserStack/local hw
End-Q: 0.0, still using Edge 15 and Safari 10 at end of quarter. (Note: Use Safari 11 in Travis builds by @csnardi landed on Jan 8.)

KR: Triage issues in a timely manner (P2)

Owner: mattl
Scoring: For all web-platform-tests infra issues, priority labels were added within 72 hours (to allow for people doing their ecosystem rotation at different times of the day or in different time zones and weekends).
Calculation: first priority label assignment datetime - pull request open datetime
Mid-Q: TODO a script is needed to calculate this/simple dashboard
End-Q: 0.7. This score is a guesstimate. The ecosystem infra rotation has been triaging these issues, and the query has tended to stay empty. Measuring the exact latency is not interesting, and since it will not be a dedicated KR in 2018 Q1, we'll just say it's been going OK - 0.7.

KR: Resolve urgent issues in a timely manner (P1)

Owner: mattl
Urgent issues:

- web-platform-tests urgent issues
- wptdashboard urgent issues
- wpt-pullresults urgent issues

Scoring: For all urgent issues, once notified mattl will work on the issue during business hours with some expectation of resolution the same day where possible (within 24 hours)
Calculation: first priority label assignment datetime - comment stating resolution datetime (closing the issue might not be the best move here, so basing it on comment datetime)
Mid-Q: TODO a script is needed to calculate this/simple dashboard Can measure time to resolve. Or something based on the above sampling.
End-Q: 7/9=0.78. There have been 3+3+3=9 urgent issues. All except Latest run upload was 2 weeks ago (Nov 14) and Build errors after no-op file change were quickly resolved.

KR: Resolve web-platform-tests infra roadmap issues (P3)

Owner: mattl, gsnedders
Add tests and documentation
Scoring: Percentage of resolved infra priority:roadmap issues opened by November 1
Mid-Q: Red: 2/45=0.04 (45 issues created before Nov 1)
End-Q: 8/45=0.18

O: web-platform-tests dashboard (wpt.fyi) is accurate and useful (P2)

KR: Consistent results for all browsers are updated every 24 hours (P1)

Owner: mattl
Rationale: Consistent means that the same commit of wpt has been run for all. Otherwise the results are harder to interpret. Should resolve On front page, by default display test runs from last complete run, and working towards Test every commit of web-platform-tests within 1 hour.
Scoring: Based on time from picking a WPT commit to that commit showing on the dashboard for all browsers. Some % below 24 hours. *Measured in the month of December*.
Mid-Q: TODO a script is needed to calculate this/simple dashboard
End-Q: 0.0. Philip's comments:

> There were no runs in November or December with results from all four browsers for the same commit. It was a mistake to combine timeliness and consistency, and in 2018 Q1 those are separate KRs. However, this not just a metrics problem. We had unclear ownership of the dashboard, significant changes landed, and an internal transition at Bocoup which all combined led to an emergency of stale results. See postmortem.

KR: Define and land a per-directory interop health metric for wpt.fyi (P1)

Owner: dknox, markdittmer
Scoring:

- +0.2 for PRD
- +0.2 for design doc with consensus
- +0.2 for implementation
- +0.2 for launch
- +0.2 for incorporating feedback from other browsers after launch

Mid-Q: Yellow: The problem is now well understood, but still need to sort out order-of-implementation of a few related wpt.fyi tasks.
End-Q: 0.5 (up to and including half of implementation -- not all of implementation landed in master)

KR: Define and land per-browser reports of tests worth prioritizing (P1)

Owner: dknox, markdittmer
Scoring: same as for previous KR
Mid-Q: Yellow: The problem is now well understood, but still need to sort out order-of-implementation of a few related wpt.fyi tasks.

This could possibly take the form of per-browser-and-directory metrics to first help identify which areas are most in need of attention, but just reports would be useful, and side-by-side comparisons would not be.

End-Q: 0.5 -- lots of great discussion on how this should look; much of the implementation is even done (but not landed)

KR: Include results from recent Edge and Safari versions (P2)

Owner: mattl
Scoring: same as corresponding KR in above objective
Mid-Q: Yellow: 0.4 as Safari 11 is included

KR: Widespread issues impacting test accuracy have been addressed (P3)

Owner: mattl
Identify root causes of unexecuted tests at https://bocoup.github.io/wpt-error-report/ and tests reported as failing in all browsers. Fix fixable causes.

Rationale: We don't know if unexecuted tests are broadly due to feature detection in tests, bugs, or errors in the test-running infrastructure.

Scoring: By November 1, identify a group of 70-100 tests that either fail in all browsers or do not execute in at least one browser. These tests can be related.

By December 31, identify and document the root causes for these failures. If the issue is a bug in the reporting tools, test runner, or upstream driver binaries or browsers, open an issue for that bug.

Score 0.01 for each test documented as either addressed by a bug report or failing by design.

Mid-Q: TODO a script is needed to calculate progress/simple dashboard
End-Q: TODO

KR: Propose time-based metrics/reports for wpt.fyi (P3)

Owner: markdittmer
Scoring:

- +0.7 for a section in WPT Dashboard Metrics
- +0.3 for a dedicated design doc about time-based metrics

Mid-Q: Yellow: Approach to other metrics is trying to account for easing the rollout of these metrics.
End-Q: 0.0

O: Make tangible progress on known real-world interop issues (P3)

This is a continuation of the "Develop a plan for improving web developer interop feedback loop" Q3 objective. The work is described in "Compat Experiment."

KR: Resolve 10 of the interop bugs affecting Google Search (P3)

Owner: foolip, kereliuk, lukebjerring, rbyers
Scoring: +0.1 for each of the interop issues that are resolved
Mid-Q: Red: 1/10 = 0.1 (we did not define “resolve”, assuming State=Done)
End-Q: 4/10 = 0.4

O: Bikeshed is the best and most reliable spec authoring tool (P2)

Owner: tabatkins

KR: Test code changes against all known Bikeshed specs (P2)

Scoring: Should resolve Regression-test against all known Bikeshed specs.

- +0.5 for adding such regression tests
- +0.2 if they are updated automatically weekly or more often
- +0.1 for each code change made where the regression tests were useful

Mid-Q: Yellow: Collecting source .bs in https://github.com/foolip/bikeshed-tests and did some work in Bikeshed to allow importing these.

KR: Support linking to tests from specs (P2)

Scoring: Should resolve Inline WPT test information into specs.

- +0.5 for landing the feature in Bikeshed
- +0.1 for each spec that links to their tests

Mid-Q: Yellow: Discussed at web standards offsite and TPAC, have a very good idea of what should be included for the initial feature.

Related OKRs

- Predictability 2017 Q2 OKRs
- Ecosystem Infra 2017 Q3 OKRs
- Ecosystem Infra 2018 OKRs
- Ecosystem Infra 2018 Q1 OKRs