# Building and Scaling a Data Stewardship Program for Products Used by Hundreds of Millions of People

Rebecca Weiss
Director of Data Science, Mozilla

2019 USENIX Conference on Privacy Engineering Practice and Respect (PEPR)

## Stages of this talk

Mozilla's mission and the business of building Firefox

Data collection for a browser company with a privacy agenda

The cost of data stewardship

Takeaways

# These are all valid questions you should be asking right now

What’s a data scientist doing talking about

data stewardship to a room full of privacy engineers?

Isn’t Mozilla all about privacy?

Doesn’t that mean you don’t collect data?

What does a data science team do at a place with no data?

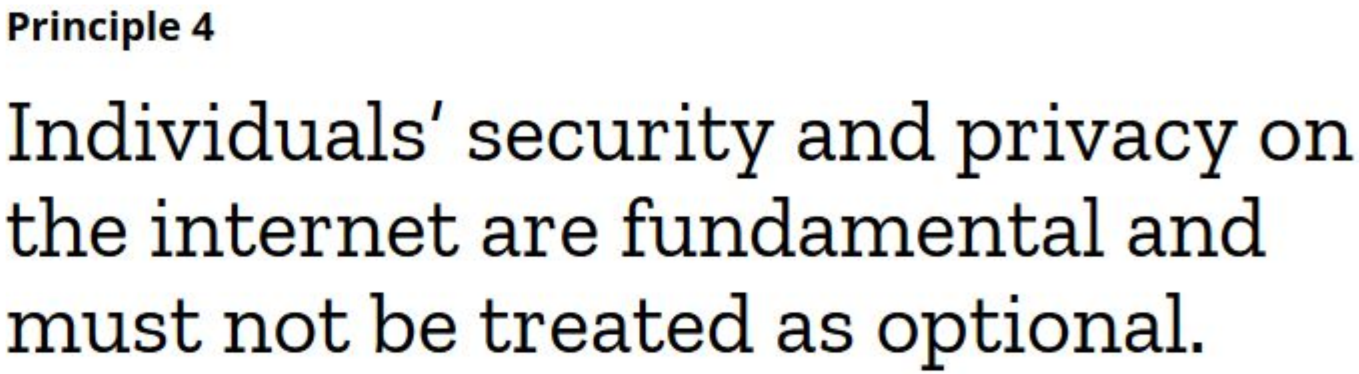
moz://a

https://data.firefox.com

Mozilla Corporation (my employer)
is a wholly owned subsidiary of
Mozilla Foundation,
a 501(c)(3) non-profit organization.

Yes, this is weird.

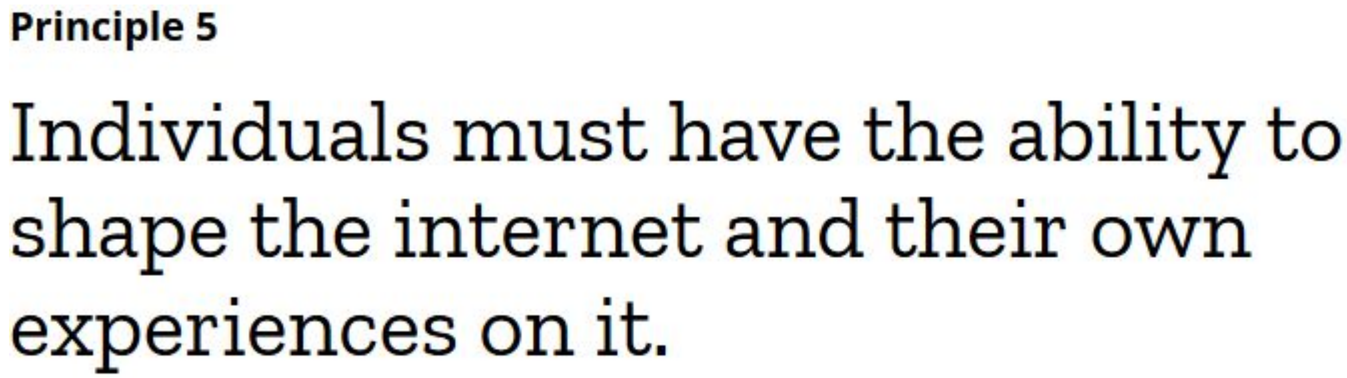
## We have a manifesto that guides us.



**Principle 4**

Individuals' security and privacy on the internet are fundamental and must not be treated as optional.

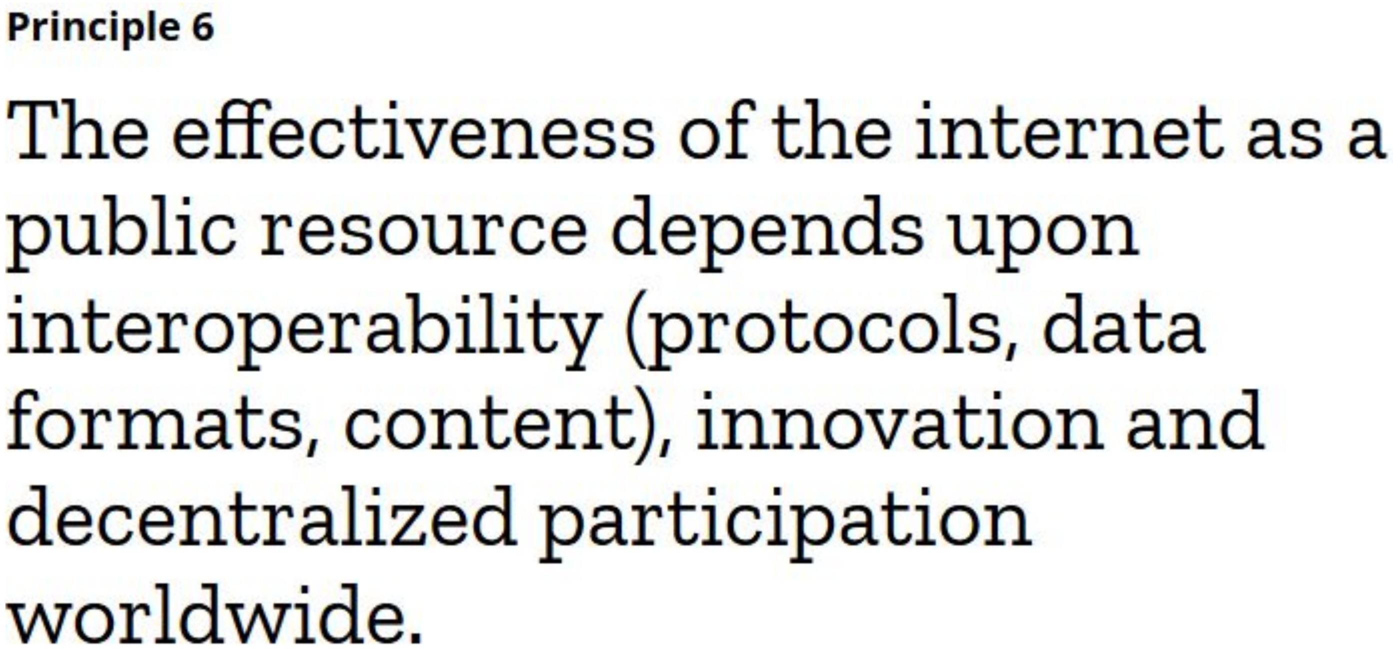
https://www.mozilla.org/en-US/about/manifesto/



**Principle 5**

Individuals must have the ability to shape the internet and their own experiences on it.

https://www.mozilla.org/en-US/about/manifesto/

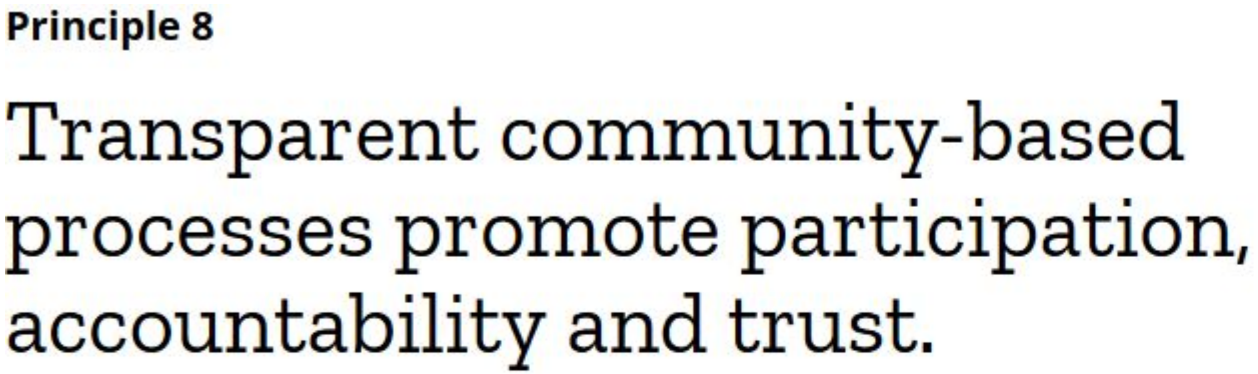


**Principle 6**

The effectiveness of the internet as a public resource depends upon interoperability (protocols, data formats, content), innovation and decentralized participation worldwide.

https://www.mozilla.org/en-US/about/manifesto/

moz://a



**Principle 8**

Transparent community-based processes promote participation, accountability and trust.

https://www.mozilla.org/en-US/about/manifesto/

# Mozilla Corporation is a taxable organization that generates revenue.

“Today, the majority of Mozilla Corporation revenue is generated from global browser search partnerships, including the deal negotiated with Google in 2017 following Mozilla’s termination of its search agreement with Yahoo/Oath which required ongoing payments to Mozilla that remain the subject of litigation.”

“In CY 2017 Mozilla Corporation generated $542 Million from royalties, subscriptions and advertising revenue compared to $506 Million in CY 2016.”

from "The State of Mozilla, 2017" (https://www.mozilla.org/en-US/foundation/annualreport/2017/)

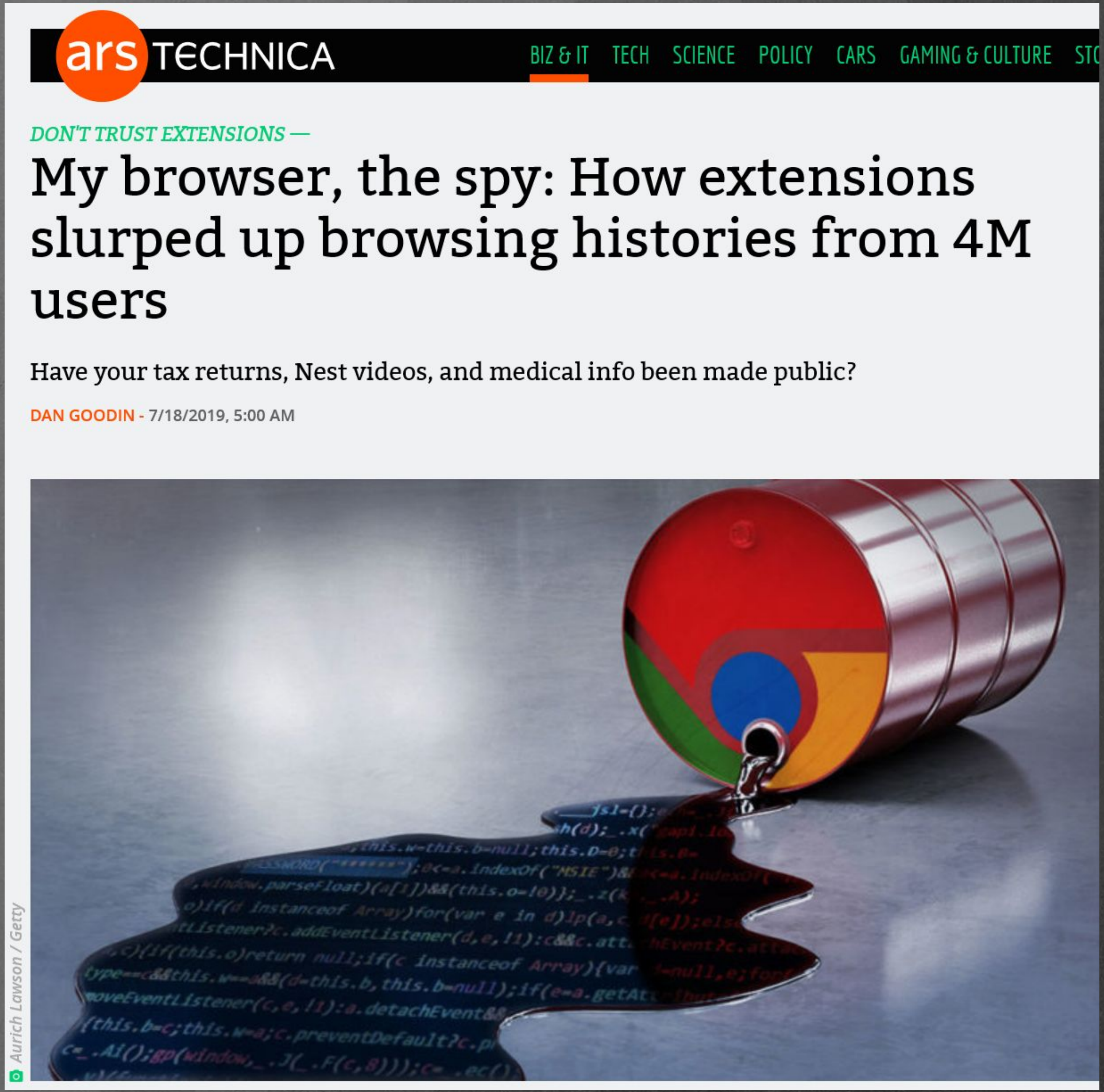
Privacy is a principle of Mozilla's mission.

Firefox enables Mozilla to exist.

Building a better Firefox requires knowing what is and isn't working well. **It is also related to our revenue.**

We collect data to enable these decisions. Collecting data is a risk to privacy.

How can we balance our commitment to privacy while also collecting data?



*DON'T TRUST EXTENSIONS —*

# My browser, the spy: How extensions slurped up browsing histories from 4M users

Have your tax returns, Nest videos, and medical info been made public?

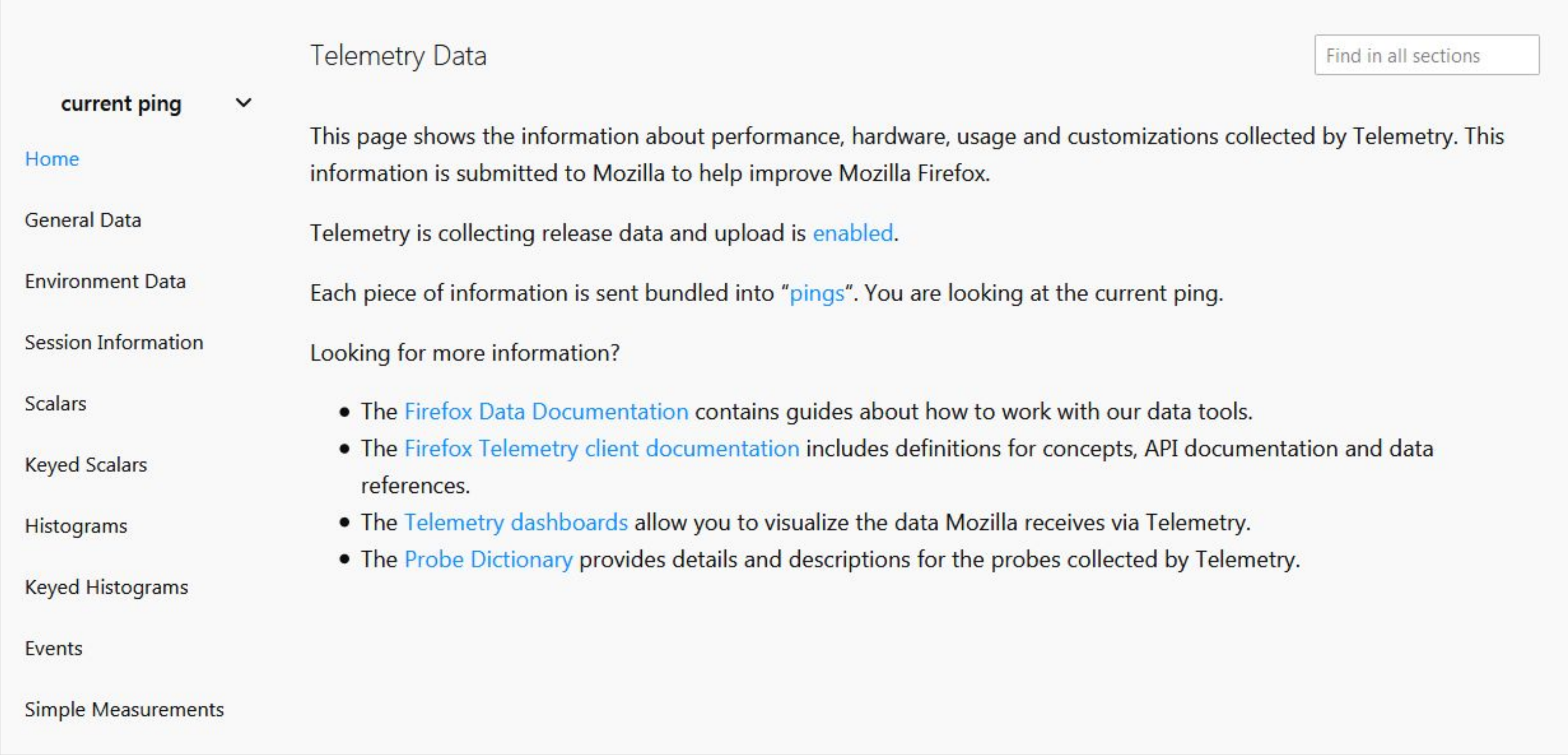
DAN GOODIN - 7/18/2019, 5:00 AM

Aurich Lawson / Getty

https://arstechnica.com/information-technology/2019/07/dataspii-inside-the-debacle-that-dished-private-data-from-apple-tesla-blue-origin-and-4m-people/

# We collect **a lot** of data.  We didn't always.

Go to **about:telemetry** in your address bar.  You can see everything that your Firefox client sends.



Telemetry Data

This page shows the information about performance, hardware, usage and customizations collected by Telemetry. This information is submitted to Mozilla to help improve Mozilla Firefox.

Telemetry is collecting release data and upload is enabled.

Each piece of information is sent bundled into "pings". You are looking at the current ping.

Looking for more information?

- The Firefox Data Documentation contains guides about how to work with our data tools.
- The Firefox Telemetry client documentation includes definitions for concepts, API documentation and data references.
- The Telemetry dashboards allow you to visualize the data Mozilla receives via Telemetry.
- The Probe Dictionary provides details and descriptions for the probes collected by Telemetry.

http://docs.telemetry.mozilla.org

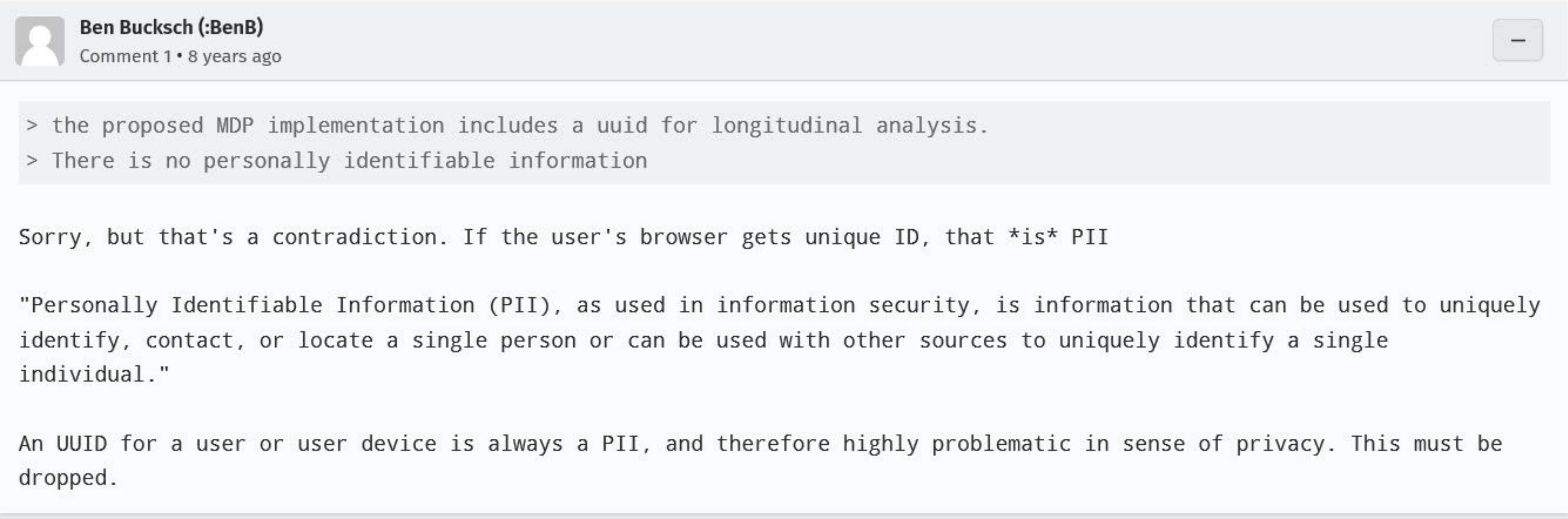
https://telemetry.mozilla.org/

https://firefox-source-docs.mozilla.org/toolkit/components/telemetry/telemetry/index.html

https://telemetry.mozilla.org/probe-dictionary/

In the beginning...

All data collection was **bad**.



**Ben Bucksch (:BenB)**

Comment 1 • 8 years ago

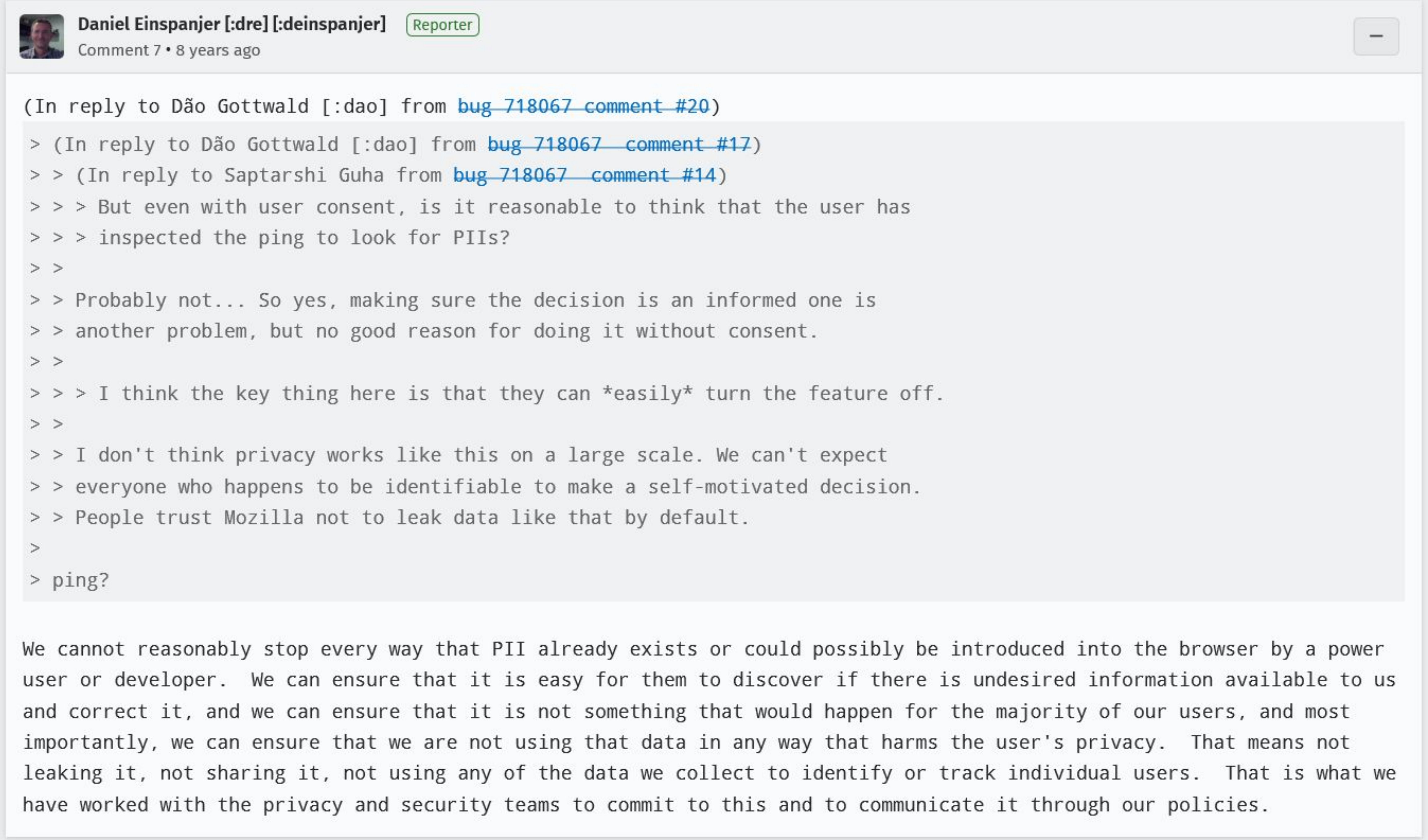
> the proposed MDP implementation includes a uuid for longitudinal analysis.
> There is no personally identifiable information

Sorry, but that's a contradiction. If the user's browser gets unique ID, that *is* PII

"Personally Identifiable Information (PII), as used in information security, is information that can be used to uniquely identify, contact, or locate a single person or can be used with other sources to uniquely identify a single individual."

An UUID for a user or user device is always a PII, and therefore highly problematic in sense of privacy. This must be dropped.

https://bugzilla.mozilla.org/show_bug.cgi?id=718066



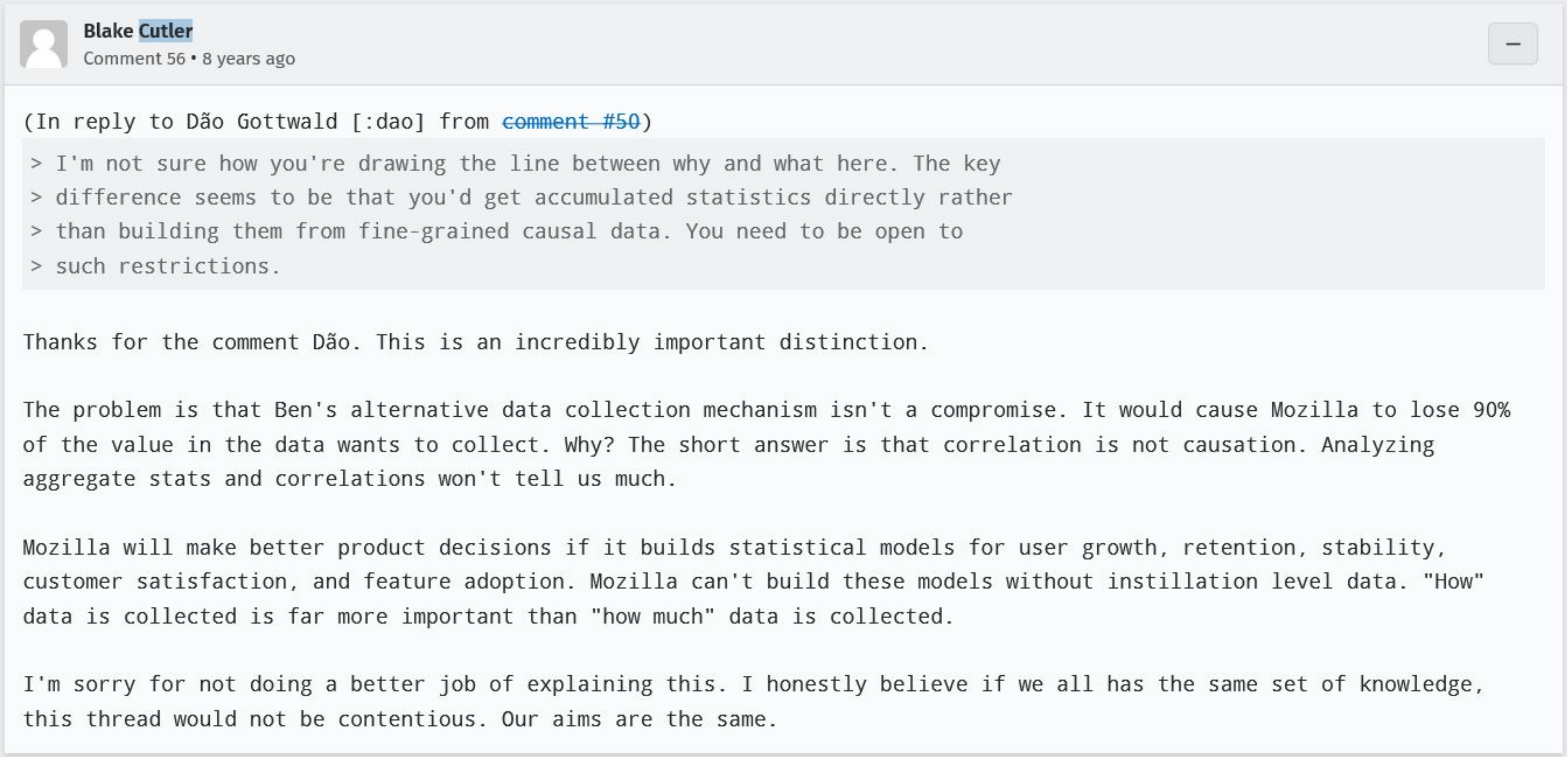
Daniel Einspanjer [:dre] [:deinspanjer] Reporter
Comment 7 • 8 years ago

(In reply to Dão Gottwald [:dao] from ~~bug 718067 comment #20~~)

> (In reply to Dão Gottwald [:dao] from ~~bug 718067 comment #17~~)
> > (In reply to Saptarshi Guha from ~~bug 718067 comment #14~~)
> > > But even with user consent, is it reasonable to think that the user has
> > > inspected the ping to look for PIIs?
> >
> > Probably not... So yes, making sure the decision is an informed one is
> > another problem, but no good reason for doing it without consent.
> >
> > > I think the key thing here is that they can *easily* turn the feature off.
> >
> > I don't think privacy works like this on a large scale. We can't expect
> > everyone who happens to be identifiable to make a self-motivated decision.
> > People trust Mozilla not to leak data like that by default.
>
> ping?

We cannot reasonably stop every way that PII already exists or could possibly be introduced into the browser by a power user or developer.  We can ensure that it is easy for them to discover if there is undesired information available to us and correct it, and we can ensure that it is not something that would happen for the majority of our users, and most importantly, we can ensure that we are not using that data in any way that harms the user's privacy.  That means not leaking it, not sharing it, not using any of the data we collect to identify or track individual users.  That is what we have worked with the privacy and security teams to commit to this and to communicate it through our policies.

https://bugzilla.mozilla.org/show_bug.cgi?id=718066



**Blake Cutler**

Comment 56 • 8 years ago

(In reply to Dão Gottwald [:dao] from ~~comment #50~~)

> I'm not sure how you're drawing the line between why and what here. The key
> difference seems to be that you'd get accumulated statistics directly rather
> than building them from fine-grained causal data. You need to be open to
> such restrictions.

Thanks for the comment Dão. This is an incredibly important distinction.

The problem is that Ben's alternative data collection mechanism isn't a compromise. It would cause Mozilla to lose 90% of the value in the data wants to collect. Why? The short answer is that correlation is not causation. Analyzing aggregate stats and correlations won't tell us much.

Mozilla will make better product decisions if it builds statistical models for user growth, retention, stability, customer satisfaction, and feature adoption. Mozilla can't build these models without instillation level data. "How" data is collected is far more important than "how much" data is collected.

I'm sorry for not doing a better job of explaining this. I honestly believe if we all has the same set of knowledge, this thread would not be contentious. Our aims are the same.

https://bugzilla.mozilla.org/show_bug.cgi?id=718066

All data collection is **bad**.

Some data collection is necessary.

Make sure the collected data provides

**direct user benefit**

Data stewards - engineer/contributors who were interested in keeping Firefox honest about browser data.

Browser data = someone has to write telemetry probes according to Firefox engineering standards

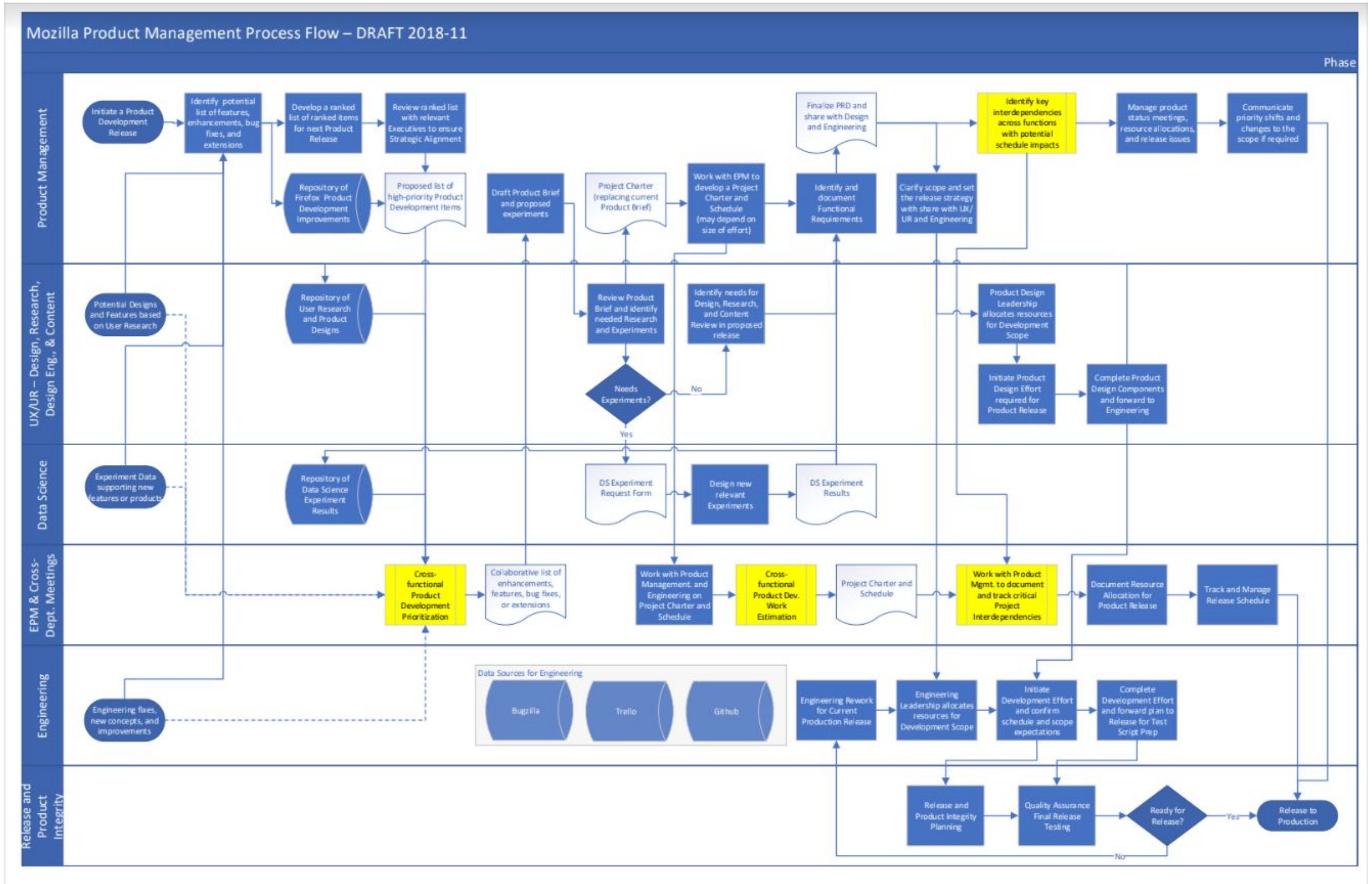
Idea: check for "direct user benefit" when you commit code with telemetry probes

One problem...this "data review" was basically subjective (and often felt adversarial)

# “Testing” for “direct user benefit” didn’t scale.

- **Scenario: Firefox feels slow when rendering performance is poor.**
- Generating the distribution of rendering performance for clients with low-power GPU requires many observations over time.
- Clients in the field behave very differently than in laboratory/testbed settings, so we need real user-generated observations to understand the problem.
- Collecting performance counters does not provide user benefit in and of themselves. It is only through follow-up analysis that user benefit is generated.
- Therefore, reviewing for user benefit became more than a test; it became a value judgment about the nature of the analysis that could be performed.
- Result: ambiguous delays in shipping software due to the methodological sophistication of the reviewer (creates incentives to get weak reviews)

# Software development lifecycle does not allow for ambiguous data review delays


Mozilla Product Management Process Flow – DRAFT 2018-11
Phase
Product Management
UX/UR – Design, Research, Design Eng., & Content
Data Science
EPM & Cross-Dept. Meetings
Engineering
Release and Product Integrity
Initiate a Product Development Release
Identify potential list of features, enhancements, bug fixes, and extensions
Develop a ranked list of ranked items for next Product Release
Review ranked list with relevant Executives to ensure Strategic Alignment
Repository of Firefox Product Development Improvements
Proposed list of high-priority Product Development Items
Draft Product Brief and proposed experiments
Project Charter (replacing current Product Brief)
Work with EPM to develop a Project Charter and Schedule (may depend on size of effort)
Finalize PRD and share with Design and Engineering
Identify and document Functional Requirements
Clarify scope and set the release strategy with share with UX/UR and Engineering
Identify key interdependencies across functions with potential schedule impacts
Manage product status meetings, resource allocations, and release issues
Communicate priority shifts and changes to the scope if required
Potential Designs and Features based on User Research
Repository of User Research and Product Designs
Review Product Brief and identify needed Research and Experiments
Identify needs for Design, Research, and Content Review in proposed release
Needs Experiments?
No
Yes
Product Design Leadership allocates resources for Development Scope
Initiate Product Design Effort required for Product Release
Complete Product Design Components and forward to Engineering
Experiment Data supporting new features or products
Repository of Data Science Experiment Results
DS Experiment Request Form
Design new relevant Experiments
DS Experiment Results
Cross-functional Product Development Prioritization
Collaborative list of enhancements, features, bug fixes, or extensions
Work with Product Management. and Engineering on Project Charter and Schedule
Cross-functional Product Dev. Work Estimation
Project Charter and Schedule
Work with Product Mgmt. to document and track critical Project Interdependencies
Document Resource Allocation for Product Release
Track and Manage Release Schedule
Engineering fixes, new concepts, and improvements
Data Sources for Engineering
Bugzilla
Trello
Github
Engineering Rework for Current Production Release
Engineering Leadership allocates resources for Development Scope
Initiate Development Effort and confirm schedule and scope expectations
Complete Development Effort and forward plan to Release for Test Script Prep
Release and Product Integrity Planning
Quality Assurance Final Release Testing
Ready for Release?
Yes
No
Release to Production

# COST

Who should bear the lion’s share of costs associated with implementing privacy?

Users or companies?

For users, that’s time and cognitive load (opportunity cost)

For companies, that’s budget spend from SOMEBODY’S cost center…

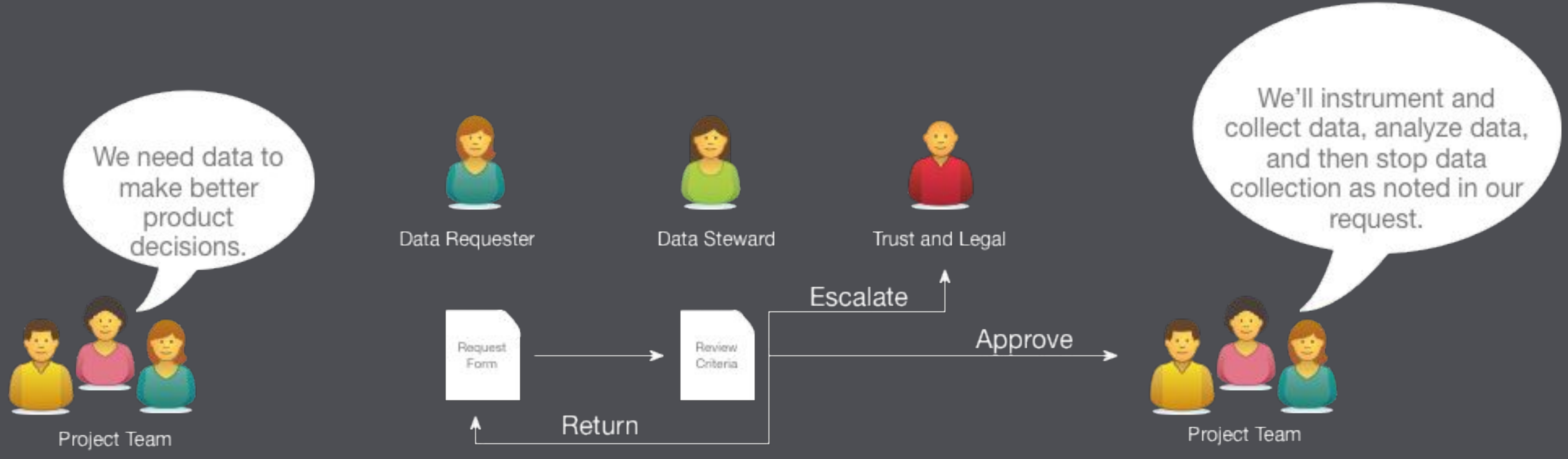
Engineering was **blowing up the cost** of privacy with a haphazard review process.

We also didn’t really have a lot of budget to spend on frivolous data that was not immediately useful.

We needed to make data reviews for transparency as close to a small, fixed cost as possible.

# Requirements for new data review process

1. We will not ask ourselves about whether you should or should not collect data.  We will only ask ourselves if this data should be opt-in or opt-out.
2. All data review will consist of the same set of questions by both requester and reviewer.  Standards for failing data review will be based entirely on the answers to these questions.
3. Paths to escalation will be clearly documented.  Escalation happens immediately when a question in data review can’t be answered or triggers the escalation response (e.g. introduction of a new UUID)



We need data to make better product decisions.

Project Team

Data Requester

Data Steward

Trust and Legal

Request Form

Review Criteria

Escalate

Approve

Return

We'll instrument and collect data, analyze data, and then stop data collection as noted in our request.

Project Team

Step 1: Complete request form and open bug

Step 2: Evalute request using review criteria.

Step 3: Escalate to trust and legal as needed.

Wait, what was that about opt-in versus opt-out?

Wait, what was that about
opt-in versus opt-out?

Let’s discuss **libertarian paternalism**.

# Cost to the user needs to be lower than the cost of employing heuristics to produce desired outcome

| User control | Cognitive load |
|---|---|
| Every preference in Firefox | High (prone to satisficing) |
| Asking for opt-in to every form of new data collection | High (prone to acquiescence bias) |
| Asking for opt-in to sensitive data collection but opt-out to data collection for analytics | Medium (libertarian paternalism) |
| Never collect data of any form | Low (💀existential threat💀) |

What does “sensitive” mean?  How do we know when we’re getting close to the red line?

# Opt-out (default on) in Release

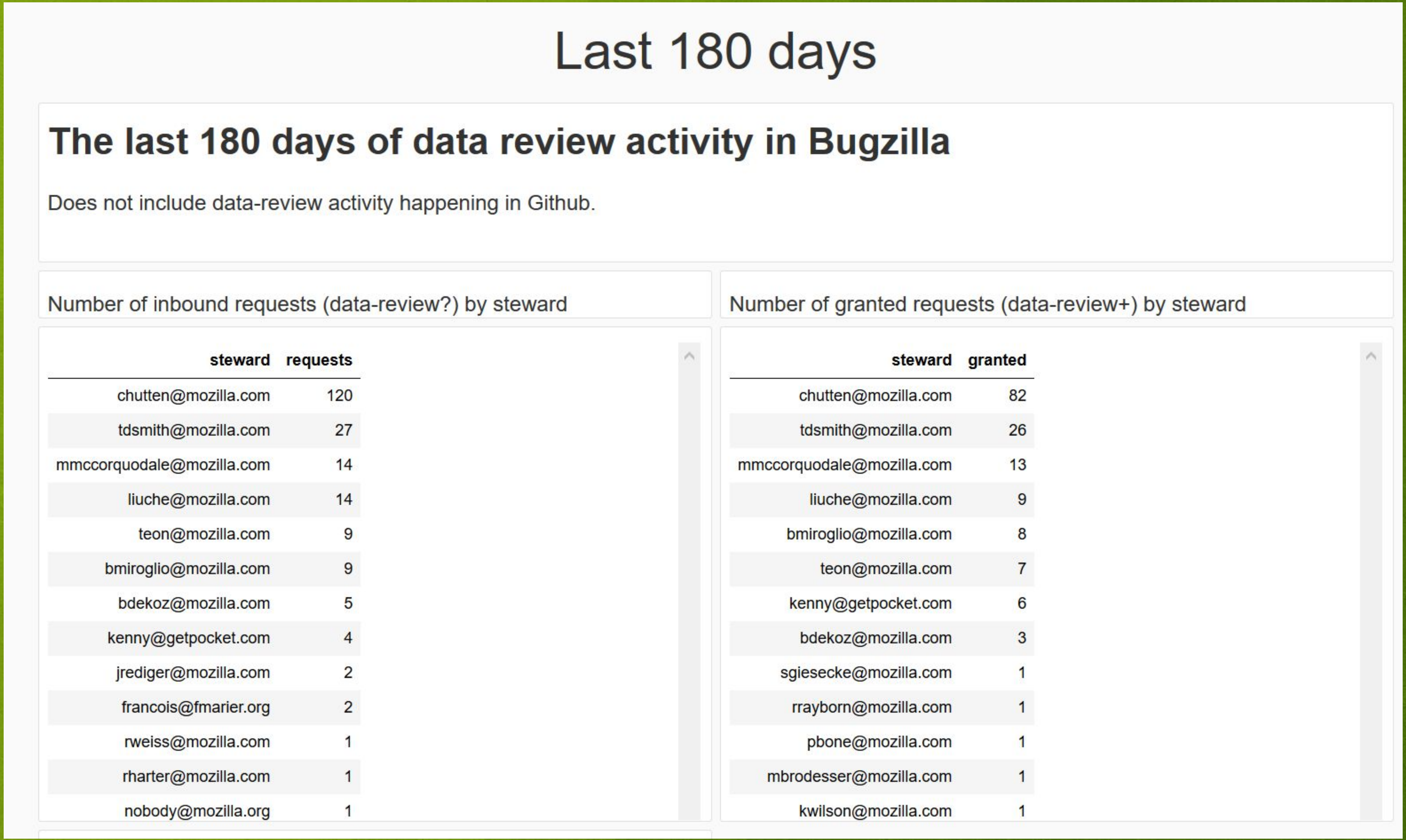
| | | | |
|---|---|---|---|
| **Category 1** | Technical information about the machine or Firefox | Examples: OS, available memory, crashes and errors, outcome of automated processes like updates, safebrowsing, activation, version, buildid, etc. | ✓ |
| **Category 1.5*** | Compatibility Information with other software | Examples: features and APIs used by websites, information about installed add-ons, or other 3rd-party software that interacts with Firefox | ✓ |
| **Category 2** | Interaction Data about user's direct engagement | Examples: number of tabs, add-ons, windows, searches via interface, use of specific features, session length, status of discrete user preferences | ✓ |

# Opt-in (default off) for Release

| | | | |
|---|---|---|---|
| **Category 3** | Web Activity Data about user browsing behavior or content | Examples: URLs of sites visited, browsing history, Interaction Data about specific pages or sites | ? |
| **Category 4** | Highly Sensitive Data that is known to be risky or personally identifiable | Examples: e-mail, usernames, identifiers such as Google Ad ID, Apple IDFA, Firefox Account identifier, saved cookies or specific website content, including memory contents, screen data, or DOM data | ✘ |

# If you're going to collect data, operate in good faith and openly by default

1. Data review is public and we are open with our motivations for data collection. We **will** collect data, but we will not hide our justification for the collection.
2. Data review checks that all data collection has meaningful control over it. If a reviewer can't find a way for the user to opt out of the data collection, it does not pass review. No r+, probe code does not land in tree.
    a. Case study: experiment with a probe counting opt-outs of Telemetry
    b. Case study: ad-clicks probe
3. Data review also cannot be passed if there is not publicly accessible documentation about what and how something is measured.



# Last 180 days

## The last 180 days of data review activity in Bugzilla

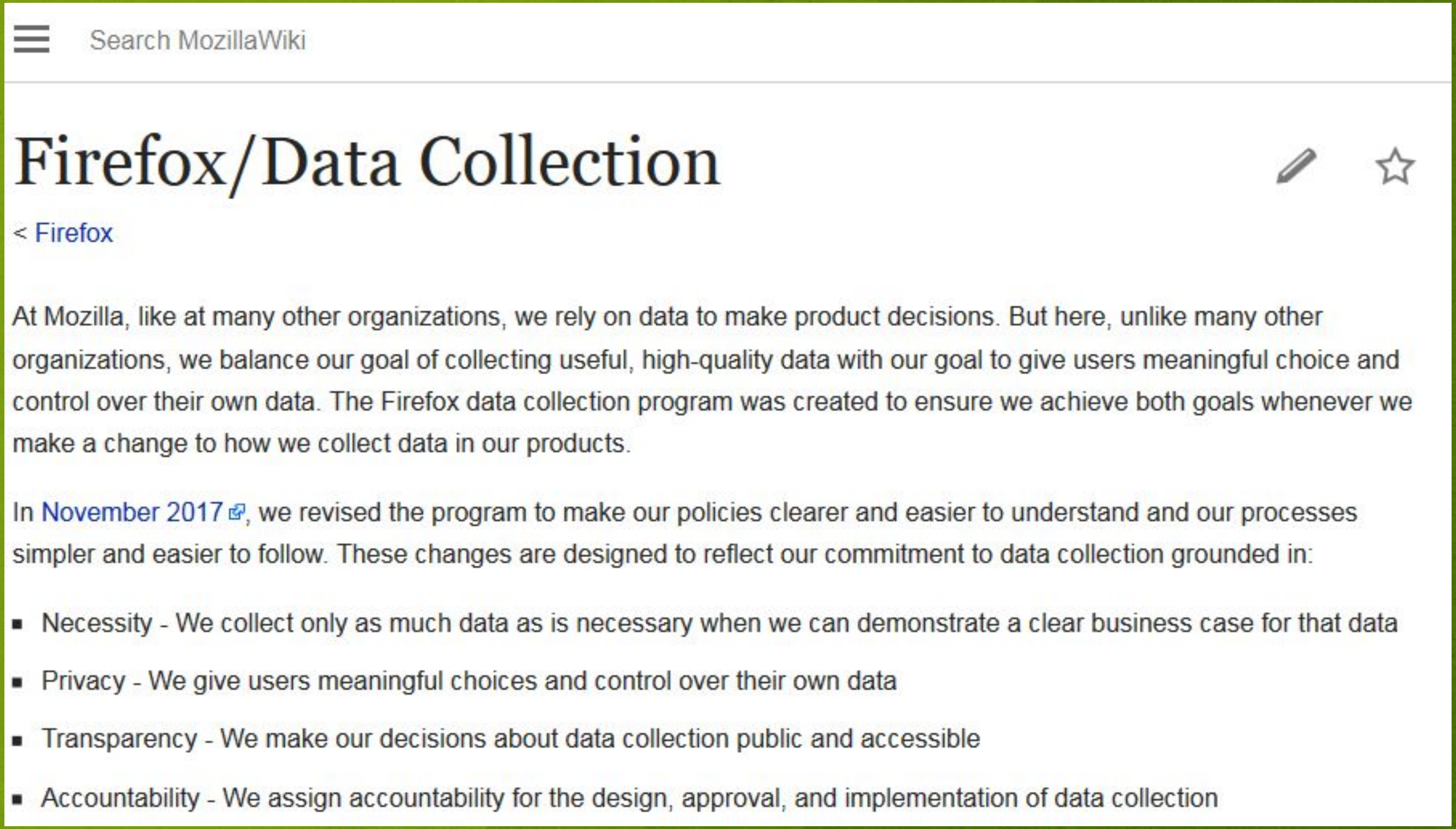
Does not include data-review activity happening in Github.

Number of inbound requests (data-review?) by steward

| steward | requests |
|---:|---:|
| chutten@mozilla.com | 120 |
| tdsmith@mozilla.com | 27 |
| mmccorquodale@mozilla.com | 14 |
| liuche@mozilla.com | 14 |
| teon@mozilla.com | 9 |
| bmiroglio@mozilla.com | 9 |
| bdekoz@mozilla.com | 5 |
| kenny@getpocket.com | 4 |
| jrediger@mozilla.com | 2 |
| francois@fmarier.org | 2 |
| rweiss@mozilla.com | 1 |
| rharter@mozilla.com | 1 |
| nobody@mozilla.org | 1 |

Number of granted requests (data-review+) by steward

| steward | granted |
|---:|---:|
| chutten@mozilla.com | 82 |
| tdsmith@mozilla.com | 26 |
| mmccorquodale@mozilla.com | 13 |
| liuche@mozilla.com | 9 |
| bmiroglio@mozilla.com | 8 |
| teon@mozilla.com | 7 |
| kenny@getpocket.com | 6 |
| bdekoz@mozilla.com | 3 |
| sgiesecke@mozilla.com | 1 |
| rrayborn@mozilla.com | 1 |
| pbone@mozilla.com | 1 |
| mbrodesser@mozilla.com | 1 |
| kwilson@mozilla.com | 1 |



# Firefox/Data Collection

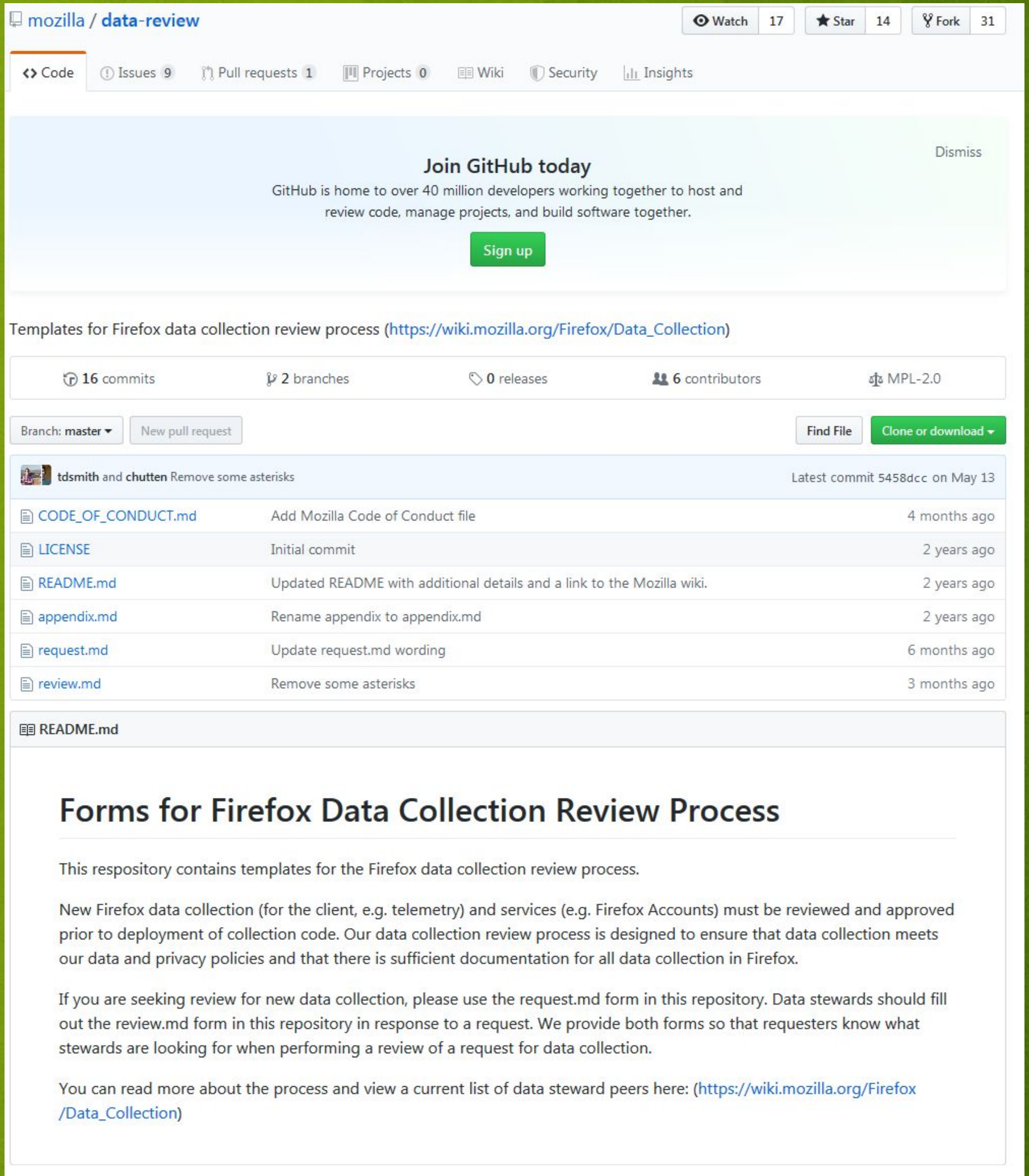
< Firefox

At Mozilla, like at many other organizations, we rely on data to make product decisions. But here, unlike many other organizations, we balance our goal of collecting useful, high-quality data with our goal to give users meaningful choice and control over their own data. The Firefox data collection program was created to ensure we achieve both goals whenever we make a change to how we collect data in our products.

In November 2017, we revised the program to make our policies clearer and easier to understand and our processes simpler and easier to follow. These changes are designed to reflect our commitment to data collection grounded in:

- Necessity - We collect only as much data as is necessary when we can demonstrate a clear business case for that data
- Privacy - We give users meaningful choices and control over their own data
- Transparency - We make our decisions about data collection public and accessible
- Accountability - We assign accountability for the design, approval, and implementation of data collection

https://wiki.mozilla.org/Firefox/Data_Collection



Join GitHub today

GitHub is home to over 40 million developers working together to host and review code, manage projects, and build software together.

Templates for Firefox data collection review process (https://wiki.mozilla.org/Firefox/Data_Collection)

tdsmith and chutten Remove some asterisks — Latest commit 5458dcc on May 13

| | | |
|---|---|---|
| CODE_OF_CONDUCT.md | Add Mozilla Code of Conduct file | 4 months ago |
| LICENSE | Initial commit | 2 years ago |
| README.md | Updated README with additional details and a link to the Mozilla wiki. | 2 years ago |
| appendix.md | Rename appendix to appendix.md | 2 years ago |
| request.md | Update request.md wording | 6 months ago |
| review.md | Remove some asterisks | 3 months ago |

README.md

# Forms for Firefox Data Collection Review Process

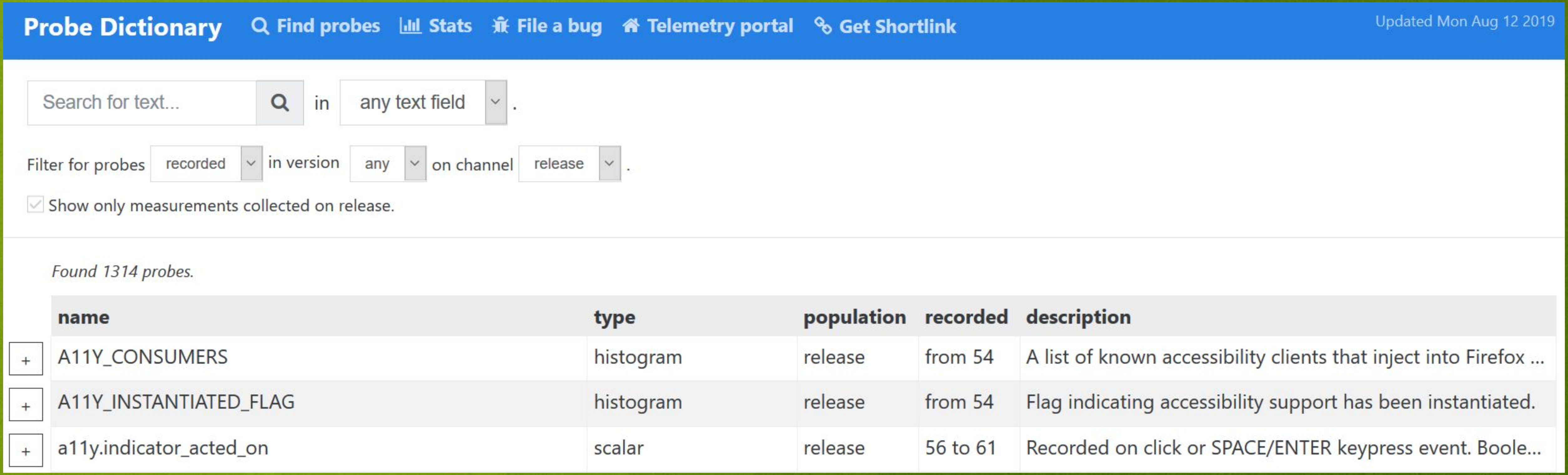
This respository contains templates for the Firefox data collection review process.

New Firefox data collection (for the client, e.g. telemetry) and services (e.g. Firefox Accounts) must be reviewed and approved prior to deployment of collection code. Our data collection review process is designed to ensure that data collection meets our data and privacy policies and that there is sufficient documentation for all data collection in Firefox.

If you are seeking review for new data collection, please use the request.md form in this repository. Data stewards should fill out the review.md form in this repository in response to a request. We provide both forms so that requesters know what stewards are looking for when performing a review of a request for data collection.

You can read more about the process and view a current list of data steward peers here: (https://wiki.mozilla.org/Firefox/Data_Collection)

https://github.com/mozilla/data-review



Search for text... in any text field .

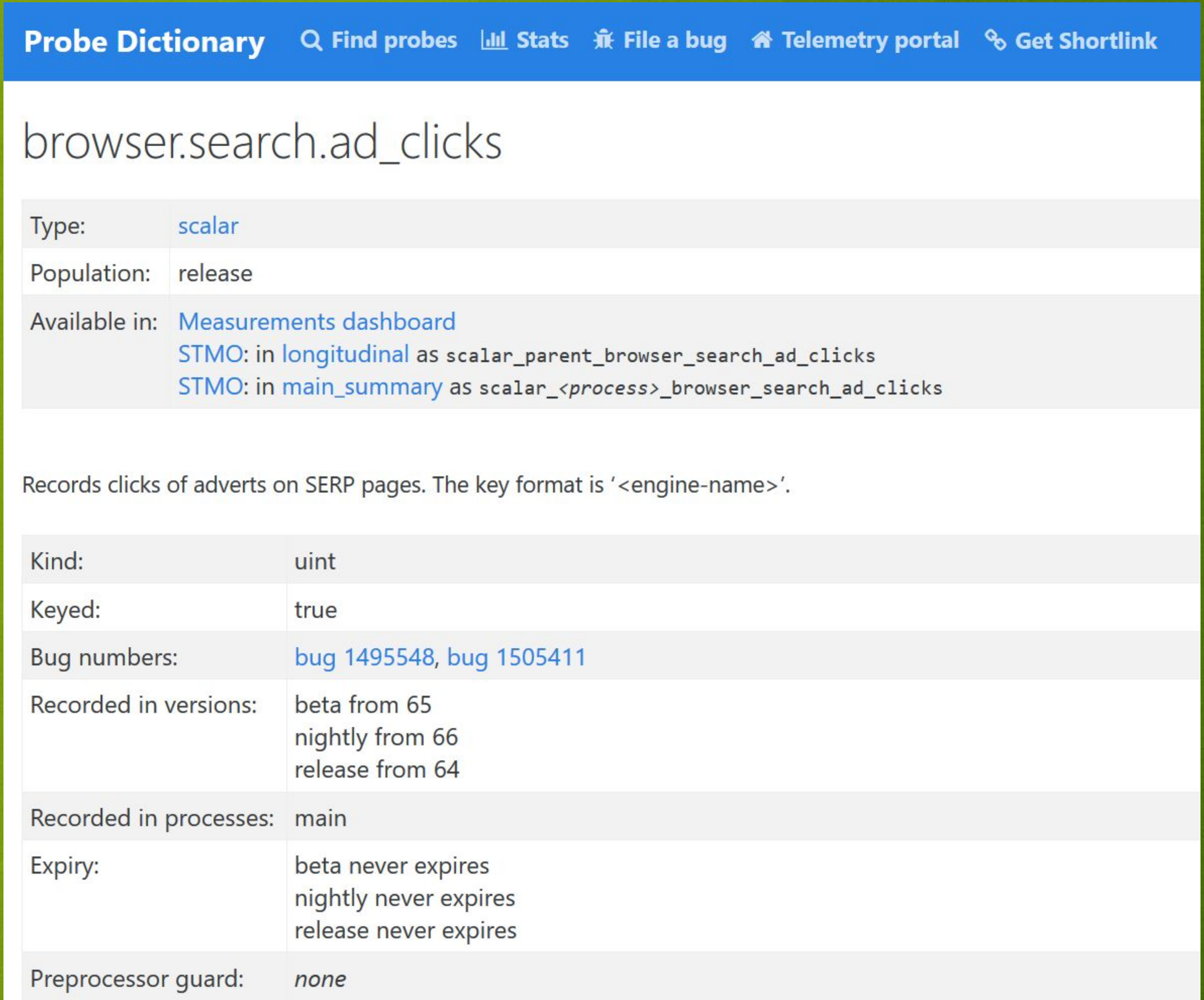
Filter for probes recorded in version any on channel release .

☑ Show only measurements collected on release.

*Found 1314 probes.*

| | name | type | population | recorded | description |
|---|---|---|---|---|---|
| + | A11Y_CONSUMERS | histogram | release | from 54 | A list of known accessibility clients that inject into Firefox ... |
| + | A11Y_INSTANTIATED_FLAG | histogram | release | from 54 | Flag indicating accessibility support has been instantiated. |
| + | a11y.indicator_acted_on | scalar | release | 56 to 61 | Recorded on click or SPACE/ENTER keypress event. Boole... |

moz://a



# browser.search.ad_clicks

| | |
|---|---|
| Type: | scalar |
| Population: | release |
| Available in: | Measurements dashboard<br>STMO: in longitudinal as `scalar_parent_browser_search_ad_clicks`<br>STMO: in main_summary as `scalar_<process>_browser_search_ad_clicks` |

Records clicks of adverts on SERP pages. The key format is ‘<engine-name>’.

| | |
|---|---|
| Kind: | uint |
| Keyed: | true |
| Bug numbers: | bug 1495548, bug 1505411 |
| Recorded in versions: | beta from 65<br>nightly from 66<br>release from 64 |
| Recorded in processes: | main |
| Expiry: | beta never expires<br>nightly never expires<br>release never expires |
| Preprocessor guard: | *none* |

https://telemetry.mozilla.org/probe-dictionary/?search=ad_clic&detailView=scalar%2Fbrowser.search.ad_clicks

# Takeaways

1. Always consider who ultimately bears the cost of implementing privacy in whatever process you create. If the user bears more of the cost, make sure they can find out why. When in doubt, increasing user agency is a safe default.
2. Product development should invest in developing governance methodologies (not just technology) as a means of mitigating privacy risk. Employees come and go, but processes and practices can still be followed when people move on.
3. High quality data is usually a good investment; good data stewardship practices lead to downstream quality improvements.
4. Your Data Science team should regularly check in with your Legal and Trust teams.
5. Fixed costs are always better than variable costs.
6. Libertarian paternalism is a slippery slope. Seek principles. Write them down. **Follow them.**
7. Transparency will (probably) save us?
8. Consider using Firefox some of the time.

moz://a

# Thank you!
# Q&A

rweiss@mozilla.com