

# Privacy and PII in relation to the web

Tomas Persson 2021-11-05

**Digitalist**

<https://digitalist.se/matomo>



matomo camp

# Who I am



<https://github.com/tomper00>

<https://www.linkedin.com/in/tomper00/>

Who we are

# Digitalist

a digital product agency

that loves Open Source

Because we know that good ideas multiply when shared

**EU**

**Sweden**



Citizen

stockholm.se

agreement

Stockholm.se  
service vendors

Hosting

development

Office 365

**Laws**

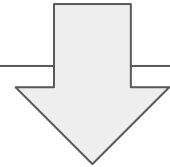
GDPR

OSL  
"Swedish  
publicity and  
secrecy law"

**USA  
(or other state)**

**Laws**

FISA 702



agreement



# What are the actual problems with Google analytics and other services?

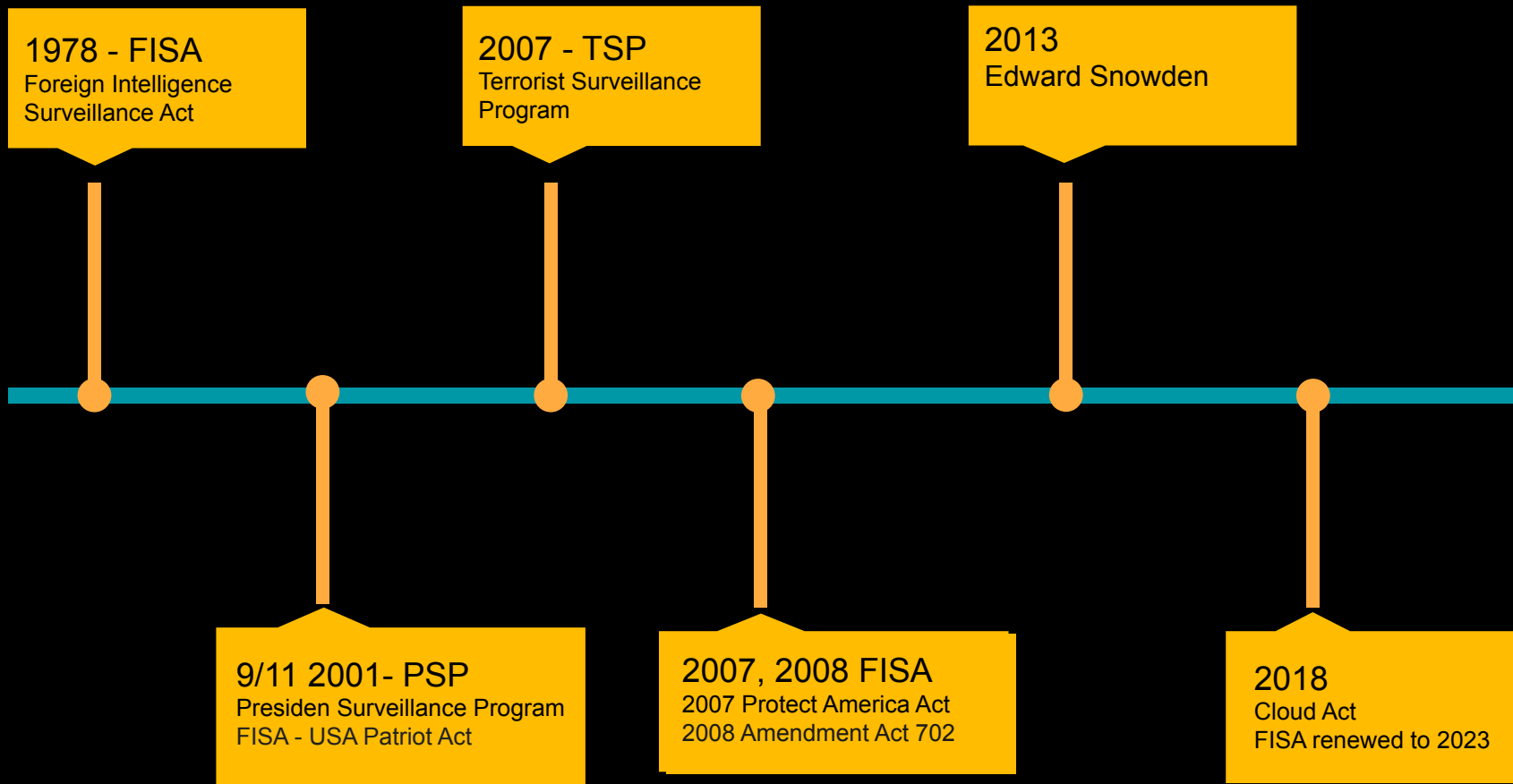
- Data leakage to foreign power
- Collection of personal data without consent
- We are selling data about our visitors



A close-up portrait of Edward Joseph Snowden, a man with short brown hair and glasses, wearing a dark suit jacket and a black shirt. He is looking slightly to the right of the camera with a neutral expression. The background is a dark, solid color.

Data leakage to foreign power

Edward Joseph Snowden





**Upstream** - eavesdropping.

**PRISM** - data collection from companies.

**TREASUREMAP** - real-time info from phones etc ..

William "Bill" Binney, former NSA technical director  
[https://youtu.be/uYg\\_0Imnr4](https://youtu.be/uYg_0Imnr4)





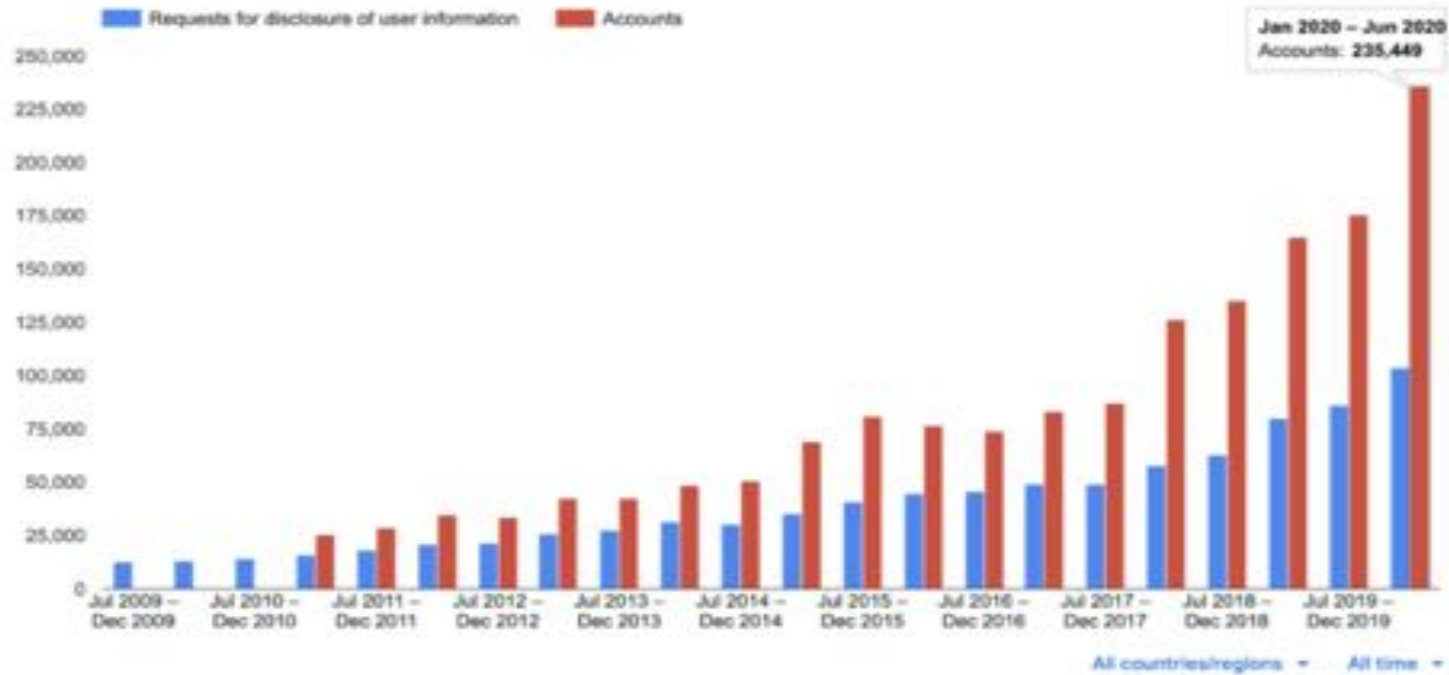
**Large amounts of internet traffic pass through the United States  
and most of it gets stuck in UPSTREAM**

**Is this really a problem in a  
country like Sweden?**

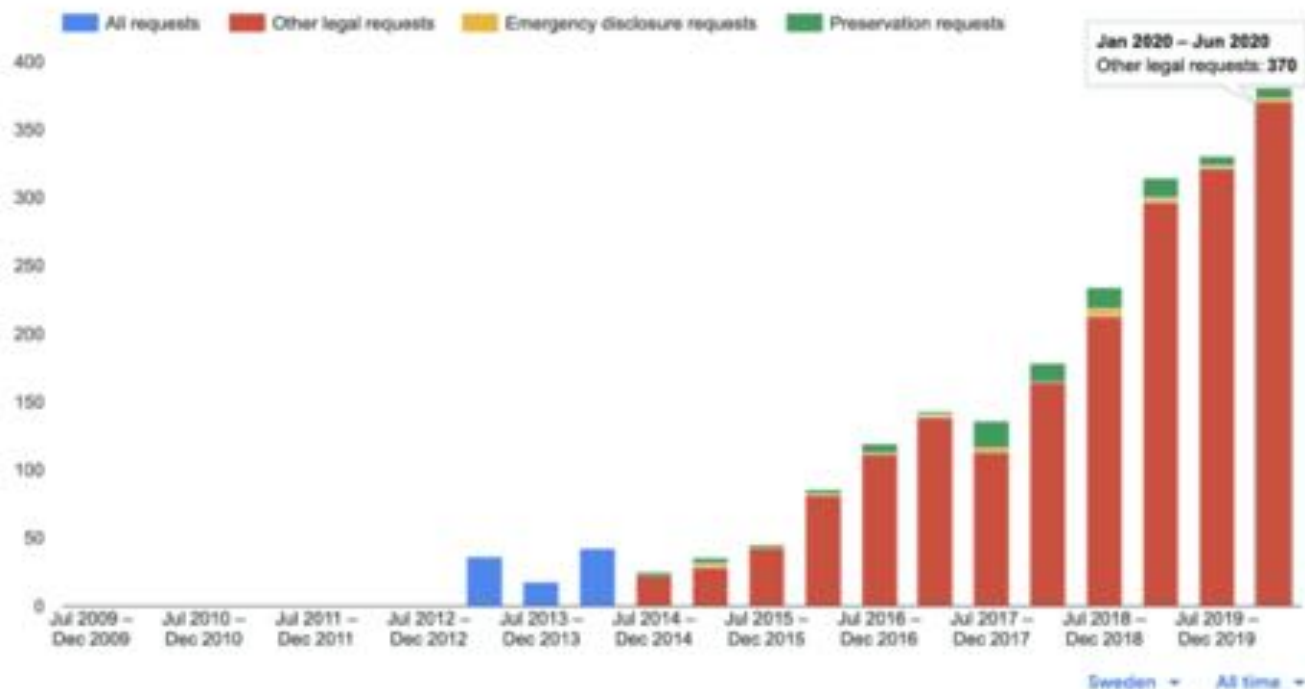
## **Privacy Reports**

**We publish this information to shine light on the impact  
government actions have on our users and the free flow  
of information online.**

# Google - Data om 235.449 användare (6 månader)



# 370 in Sweden in the same period



# Summary last half of 2019 in Sweden

Google - **321** - requests

<--- Thus increased to 370 first half year (2021)

Microsoft - **396** - requests

Apple - **92** - units

Facebook - **548** - accounts

≈ **8 cases / day** in Sweden

<https://www.microsoft.com/en-us/corporate-responsibility/law-enforcement-requests-report>

<https://www.apple.com/legal/transparency/>

<https://govtrequests.facebook.com/>

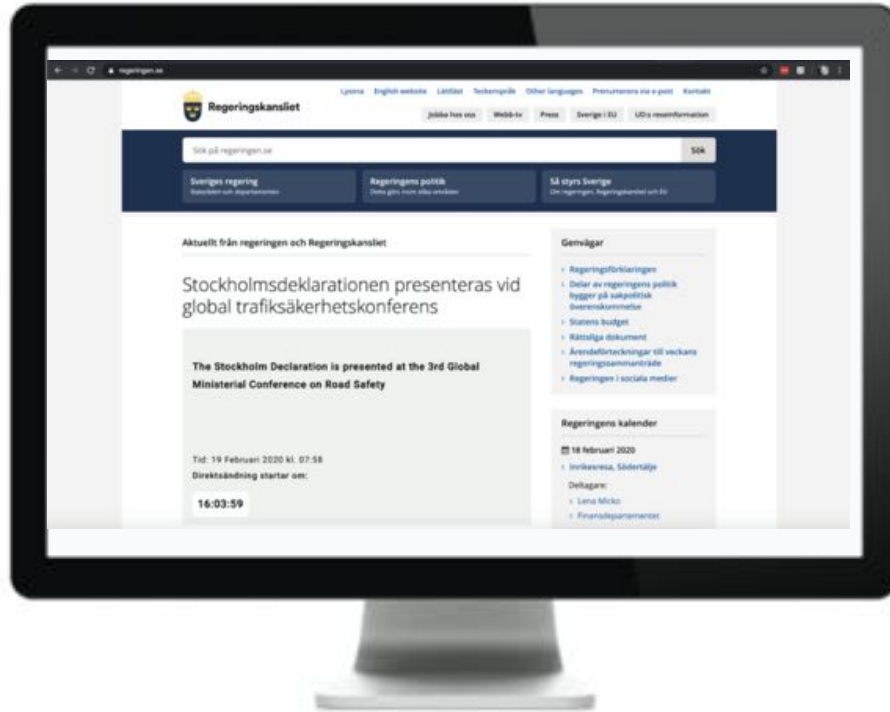
<https://transparencyreport.google.com/user-data/overview?hl=en>

# What does the public sector in Europe need to do?

- Secure the use of services (open technologies)
- Secure where the services run and who has access to them (secure cloud services)
- Secure traffic between residents and authorities (so that it never leaves Europe)

# **Leakage of personal data online web**

# What happens on the web when we visit 5 sites?







# Remember - Google sells info about our visitors

83%



**But Google Analytics is only a small  
part of the bigger problem**

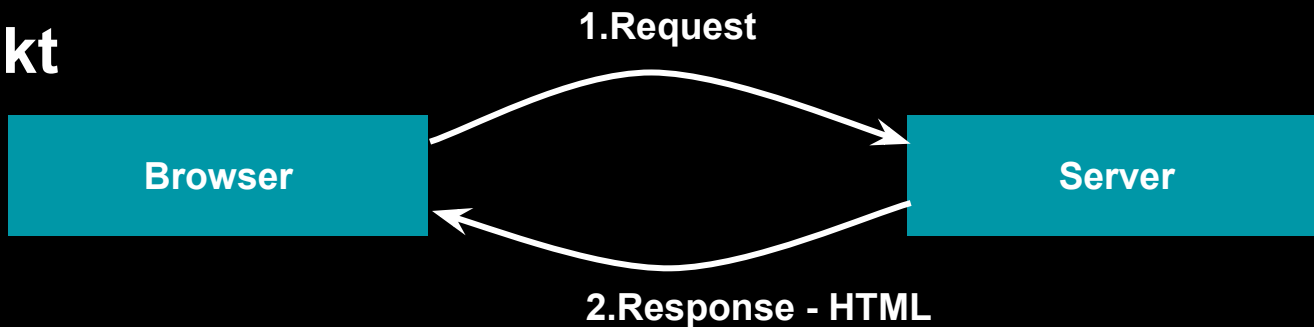
🔒 [regeringen.se/kontakt/](https://regeringen.se/kontakt/)

# A technical deep dive at the Swedish Gov



The screenshot displays the top section of the Swedish Government website. On the left is the logo of the Regeringskansliet (Government Offices), featuring a crown and the text "Regeringskansliet". To the right of the logo are several navigation links: "Lyssna", "English website", "Lättläst", "Teckenspråk", "Öther languages", "Prenumerera via e-post", and a blue "Kontakt" button. Below these links is a row of five buttons: "jobba hos oss", "Webb-tv", "Press", "Sverige i EU", and "UD:s reseinformation". A large search bar is positioned below the navigation, with the placeholder text "Sök på regeringen.se" and a "Sök" button on the right. At the bottom of the header area, there are three dark blue boxes with white text: "Sveriges regering" (with the subtitle "Statsråden och departementen"), "Regeringens politik" (with the subtitle "Detta gills inom olika områden"), and "Så styrs Sverige" (with the subtitle "Om regeringen, Regeringskansliet och EU").

# /kontakt



33 request to 6 domains

kontakt/	200	www.regeringen.se	document
modernizr.custom_min_30492.js	200	www.regeringen.se	script
bundle?v=pp8MLSOU6gyQRqs8xcM3L7kuA0_vXDN=rMGfVNY0_p01	200	www.regeringen.se	stylesheet
GoBrain.min.js	200	play2.qbrick.com	script
logo-sv.png	200	www.regeringen.se	png
find.js	200	dl.episerver.net	script
js?v=LWrcM8g6WfhwT3dPOO3DTi0Pssi88km80_hPGyHBx1g1	200	www.regeringen.se	script
ReadSpeaker.js?pids=embhl	200	www.regeringen.se	script
ajax-loader.gif	200	www.regeringen.se	gif
gtm.js?id=GTM-P4L58V	200	www.googletagmanager.com	script
opensans-regular-webfont.woff2	200	www.regeringen.se	font
opensans-sembold-webfont.woff2	200	www.regeringen.se	font
opensans-light-webfont.woff2	200	www.regeringen.se	font

## Request

Request URL: <https://dl.episerver.net/13.2.5/epi-util/find.js>

Request Method: GET

Status Code: ● 200

Remote Address: 104.18.18.118:443

Referrer Policy: no-referrer-when-downgrade

kontakt/	www.regeringen.se
modernizr.custom.min.30492.js	www.regeringen.se
bundle?v=p8MLSOU6gyQRqs8xcM3L7kuA0_vXDN-nMGfVNY0_p01	www.regeringen.se
GoBrain.min.js	play2.qbrick.com
logo-sv.png	www.regeringen.se
find.js	dl.episerver.net
js?v=LWrcM8g6WfhwT3dPOO3DTi0Pssi88km80_hPGyHBx1g1	www.regeringen.se
ReadSpeaker.js?pids=embhl	www.regeringen.se
ajax-loader.gif	
gtm.js?id=GTM-P4L58V	
opensans-regular-webfont.woff2	
opensans-semibold-webfont.woff2	
opensans-light-webfont.woff2	

## Request headers

```
:authority: dl.episerver.net
:method: GET
:path: /13.2.5/epi-util/find.js
:scheme: https
accept: */*
accept-encoding: gzip, deflate, br
accept-language: sv-SE,sv;q=0.9,en-US;q=0.8,en;q=0.7
cache-control: no-cache
cookie: __cfduid=d9688888259e51a1e4d0cf79cc34596c6157
pragma: no-cache
referrer: https://www.regeringen.se/kontakt/
sec-fetch-mode: no-cors
sec-fetch-site: cross-site
user-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10
```

The information shared in headers means that all servers we call can do the same type of analysis that we have in Google Analytics by default. (Pageviews with segmentation on technologies, IP addresses etc.)



8 calls (4 domains) reach the United States

- [google-analytics.com](https://www.google.com/analytics/)
- [google-tagmanager.com](https://www.google.com/tagmanager/)
- [stats.g.doubleclick.net](https://stats.g.doubleclick.net)
- [dl.episerver.net](https://dl.episerver.net)

In these calls, we (almost) always leak

- IP addresses
- URL

Number of calls in Sweden  
26 - [governments.se](https://www.government.se)  
1 - [play2.qbrick.com](https://play2.qbrick.com)

All calls that cross our borders basically mean that we can not live up to the GDPR and that we have no idea who has access to the data.

**A lot of other services actually have  
the same issues.**



## Other commonly used scripts with the same problem:

- Readspeaker
- SiteImprove statistics
- Apsis
- EpiServer find
- NewRelic
- etc

**PII**

**(Personally identifiable information)**

- IP addresses
- Email addresses
- ZIP code
- Social security number
- Credit card number
- Name
- Phone number
- etc.



Personal data



Personal data is also  
Data that in  
combination with  
other data can  
identify a person

# When anonymous data is no longer anonymous

Metadata in our web analysis

Lomträsk + smith

Lomträsk + turkish + woman

Lomträsk + homosexual

Lomträsk + unusual mobile phone



**Lomträsk**  
**55 inhabitants**

erige

Norge

Finland

**Lets privacy test your site**

<https://github.com/tomper00/privacy-test-your-site>

**What do we do when  
the accident occurs?**



**There must always be routines in place  
that assume that we will find PII.**



# PII incident

## Example of action Plan

1. Assess severity
2. Prevent further collection
3. Decide on any report to your data protection authority (within 72 hours)
4. Clear data
5. Document incidents
6. Inform victims?
7. Retrospective

## Deleting data in Matomo

1. Find the data
2. Find the visitor id's
3. Delete the visits
4. Reprocess the visitor log so that aggregated data is updated

# Simple example of a simple PII monitoring with the Alert Plugin

The screenshot shows the 'Manage Custom Alerts' interface. On the left, there is a table with the following data:

Name	Website	Period	Report
Find PII in urls	example.com	Day	Page URLs
PII on Event actions	example.com	Day	Event Actions

Below the table are two buttons: 'CREATE NEW ALERT' (green) and 'HISTORY OF TRIGGERED ALERTS' (grey). A yellow arrow points from the 'CREATE NEW ALERT' button to the configuration panel on the right.

The configuration panel on the right is titled 'Alert Condition' and contains the following settings:

- This applies to report: Page URLs
- when Page URL matches regular expression (Value: @[\d\d\d\d-\d\d\d-\d\d\d\d])
- Alert me when Unique Pageviews (Value: 0)
- is greater than (Value: 0)

Example regex:

Finding email-addresses:

```
^[^s@]+@[^s@]+\.[^s@]
```

Finding dates

```
[$|\d\d\d\d-\d\d\d-\d\d\d]
```

Combine multiple checks with |

```
^[^s@]+@[^s@]+\.[^s@]|[$|\d\d\d\d-\d\d\d-\d\d\d]
```

**Drawback:**

Alarms comes in (the next day)

Hard to detect things like a name in negex

# Deleting data with GDPR Tools or the API

Search for visitors (but very limited) since we will not be able to find users matching an event name containing [tomas@digitalist.se](mailto:tomas@digitalist.se)

The found results include all visits without any time restriction and include today.

✓	Site	Visit ID	Visitor ID	Visitor IP	User ID	Info	Action
✓	Digitalist.se	346239	25551191a3e614a4	83.254.0.0			<a href="#">View visitor profile</a>

[EXPORT SELECTED VISITS](#) [DELETE SELECTED VISITS](#)

## Deleting a lot of data

[https://matomo-server/index.php?module=API&method=PrivacyManager.deleteDataSubjects&visits\[0\]\[idsite\]=1&visits\[0\]\[idvisit\]=12345&token\\_auth=API\\_TOKEN](https://matomo-server/index.php?module=API&method=PrivacyManager.deleteDataSubjects&visits[0][idsite]=1&visits[0][idvisit]=12345&token_auth=API_TOKEN)

# Future ideas - create a PII monitoring plugin

Set up rules to find PII (on a wider scale) on Events, Pageviews, referrer urls etc

Set up alarms

Make it possible to add exceptions (when we find things that are ok)

## Make it easy to:

- Delete data (both in visitor logs and aggregates)
- Anonymize data (both in visitor logs and aggregates)

## 6 steps to secure your web analysis

1. Collect consents and inform visitors
2. Check your data collection (avoid collection PII) and set up Matomo properly
3. Manage the quality on your applications and websites
4. Secure your technical infrastructure for your system and control who has access
5. Limit time for how long data is stored
6. Processes in place to handle incidents and to monitor your data

# Questions