# WASP: Web Archiving and Search Personalized

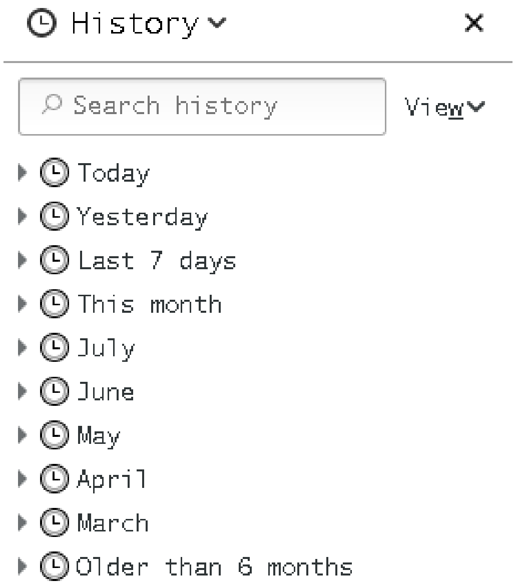
**Johannes Kiesel**, Arjen P. de Vries, Matthias Hagen, Benno Stein and Martin Potthast
@KieselJohannes, @arjenpdevries, @matthias_hagen, @bennostein, @martinpotthast

DESIRES, August 29th 2018

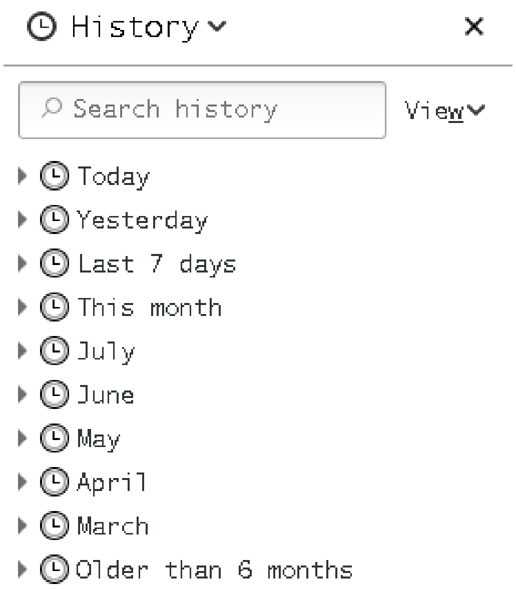
# The Personal Search Engine: Motivation

# The Personal Search Engine: Motivation

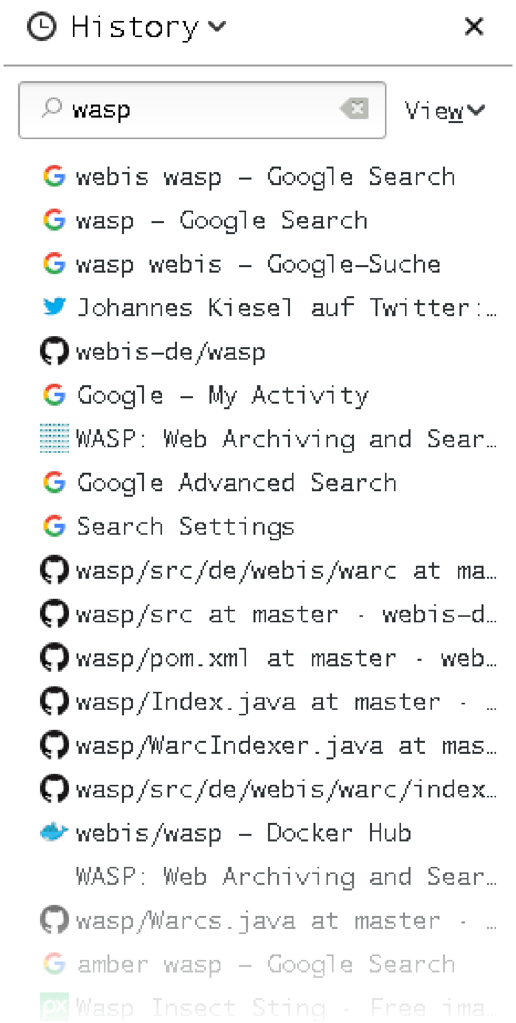
# The Personal Search Engine: Motivation



# The Personal Search Engine: Motivation



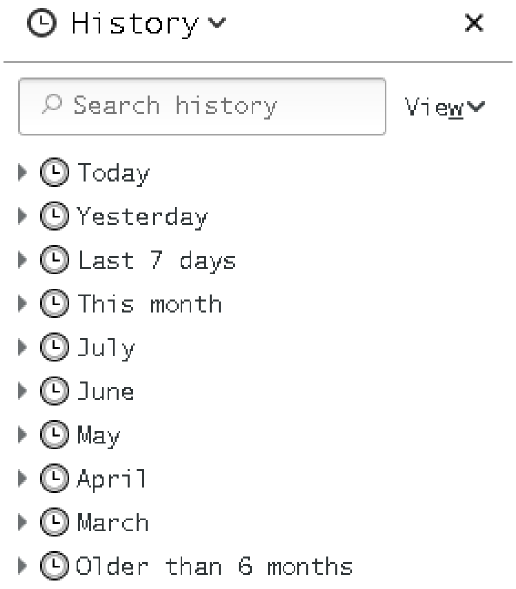
- Today
- Yesterday
- Last 7 days
- This month
- July
- June
- May
- April
- March
- Older than 6 months



wasp

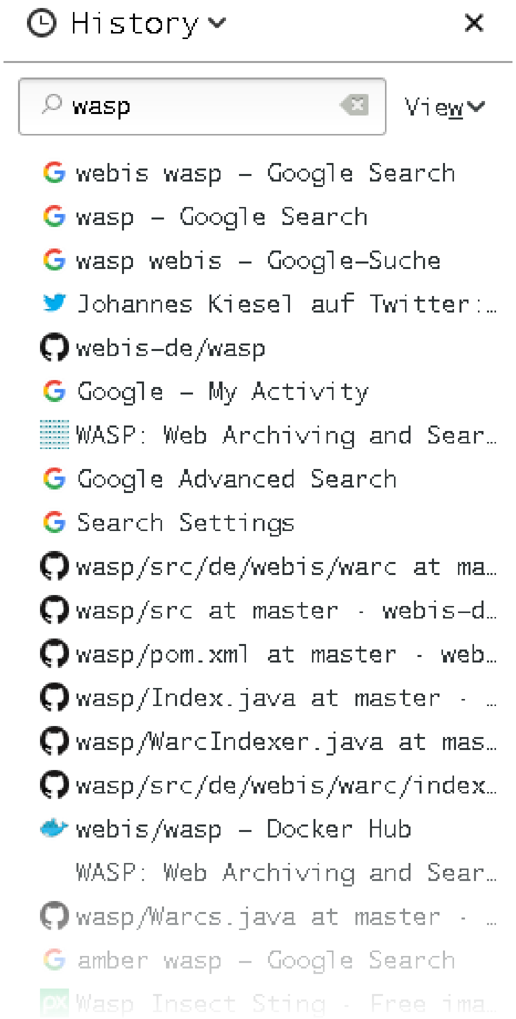
- webis wasp – Google Search
- wasp – Google Search
- wasp webis – Google-Suche
- Johannes Kiesel auf Twitter:…
- webis-de/wasp
- Google – My Activity
- WASP: Web Archiving and Sear…
- Google Advanced Search
- Search Settings
- wasp/src/de/webis/warc at ma…
- wasp/src at master · webis-d…
- wasp/pom.xml at master · web…
- wasp/Index.java at master · …
- wasp/WarcIndexer.java at mas…
- wasp/src/de/webis/warc/index…
- webis/wasp – Docker Hub
- WASP: Web Archiving and Sear…
- wasp/Warcs.java at master · …
- amber wasp – Google Search
- Wasp Insect Sting · Free ima…

# The Personal Search Engine: Motivation



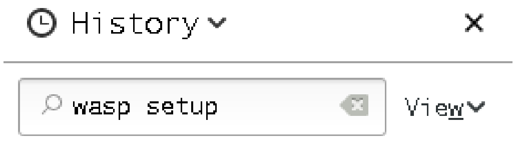
Search history

- Today
- Yesterday
- Last 7 days
- This month
- July
- June
- May
- April
- March
- Older than 6 months



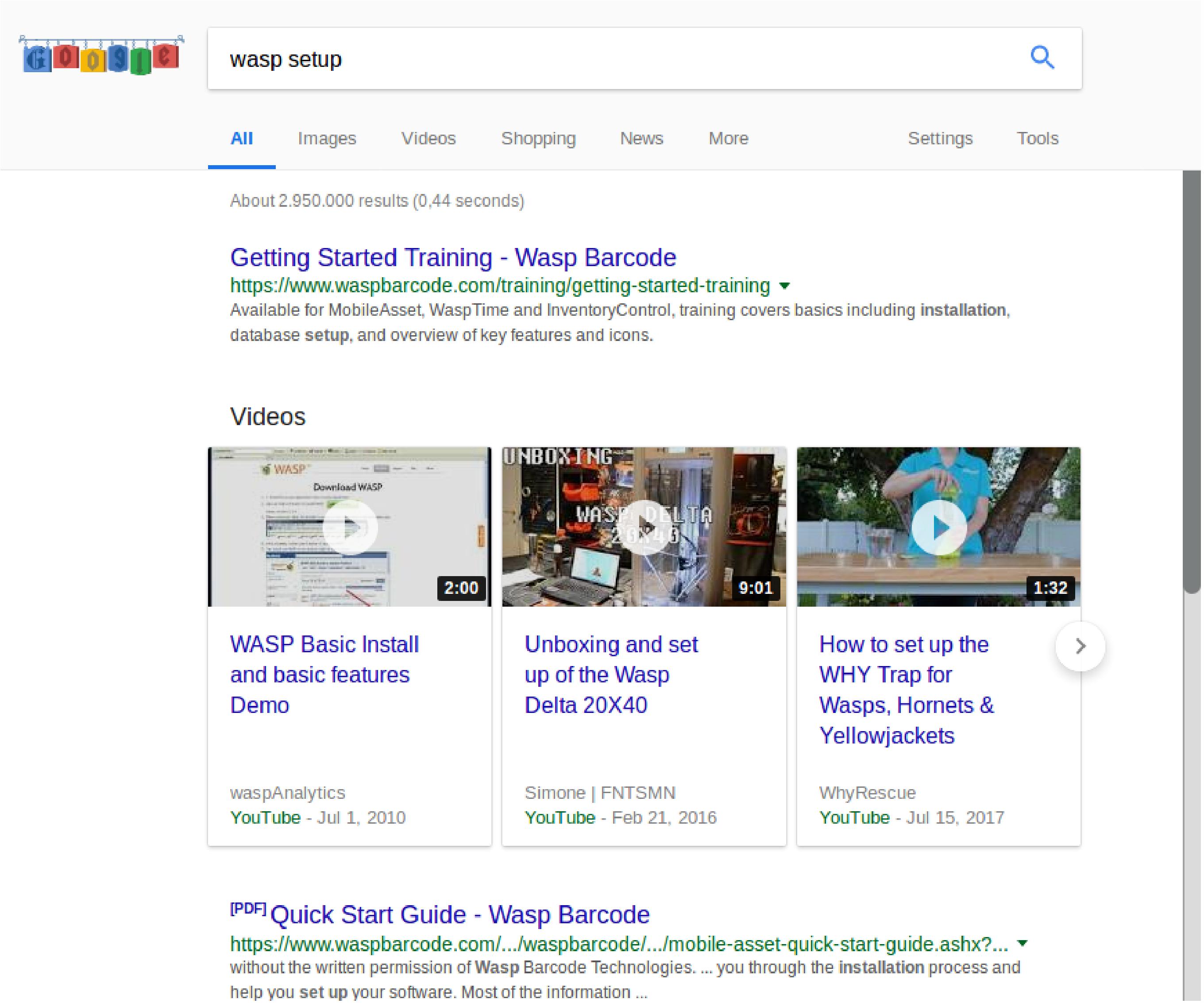
wasp

- webis wasp – Google Search
- wasp – Google Search
- wasp webis – Google-Suche
- Johannes Kiesel auf Twitter:…
- webis-de/wasp
- Google – My Activity
- WASP: Web Archiving and Sear…
- Google Advanced Search
- Search Settings
- wasp/src/de/webis/warc at ma…
- wasp/src at master · webis-d…
- wasp/pom.xml at master · web…
- wasp/Index.java at master · …
- wasp/WarcIndexer.java at mas…
- wasp/src/de/webis/warc/index…
- webis/wasp – Docker Hub
- WASP: Web Archiving and Sear…
- wasp/Warcs.java at master · …
- amber wasp – Google Search
- Wasp Insect Sting · Free ima…



wasp setup

# The Personal Search Engine: Motivation



About 2.950.000 results (0,44 seconds)

Getting Started Training - Wasp Barcode
https://www.waspbarcode.com/training/getting-started-training
Available for MobileAsset, WaspTime and InventoryControl, training covers basics including **installation**, database **setup**, and overview of key features and icons.

Videos

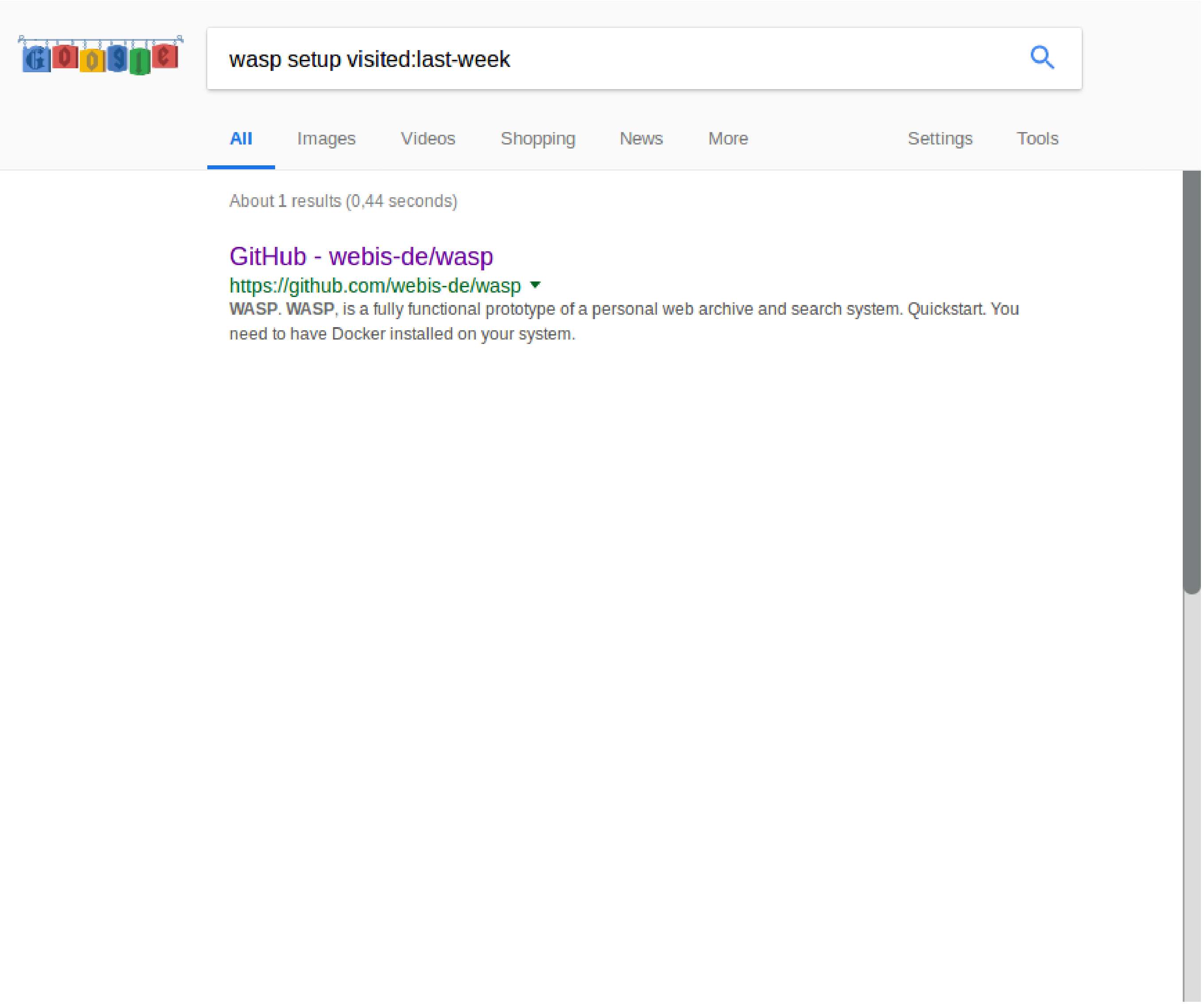
WASP Basic Install and basic features Demo
waspAnalytics
YouTube - Jul 1, 2010
2:00

Unboxing and set up of the Wasp Delta 20X40
Simone | FNTSMN
YouTube - Feb 21, 2016
9:01

How to set up the WHY Trap for Wasps, Hornets & Yellowjackets
WhyRescue
YouTube - Jul 15, 2017
1:32

[PDF] Quick Start Guide - Wasp Barcode
https://www.waspbarcode.com/.../waspbarcode/.../mobile-asset-quick-start-guide.ashx?...
without the written permission of **Wasp** Barcode Technologies. ... you through the **installation** process and help you **set up** your software. Most of the information ...

# The Personal Search Engine: Motivation

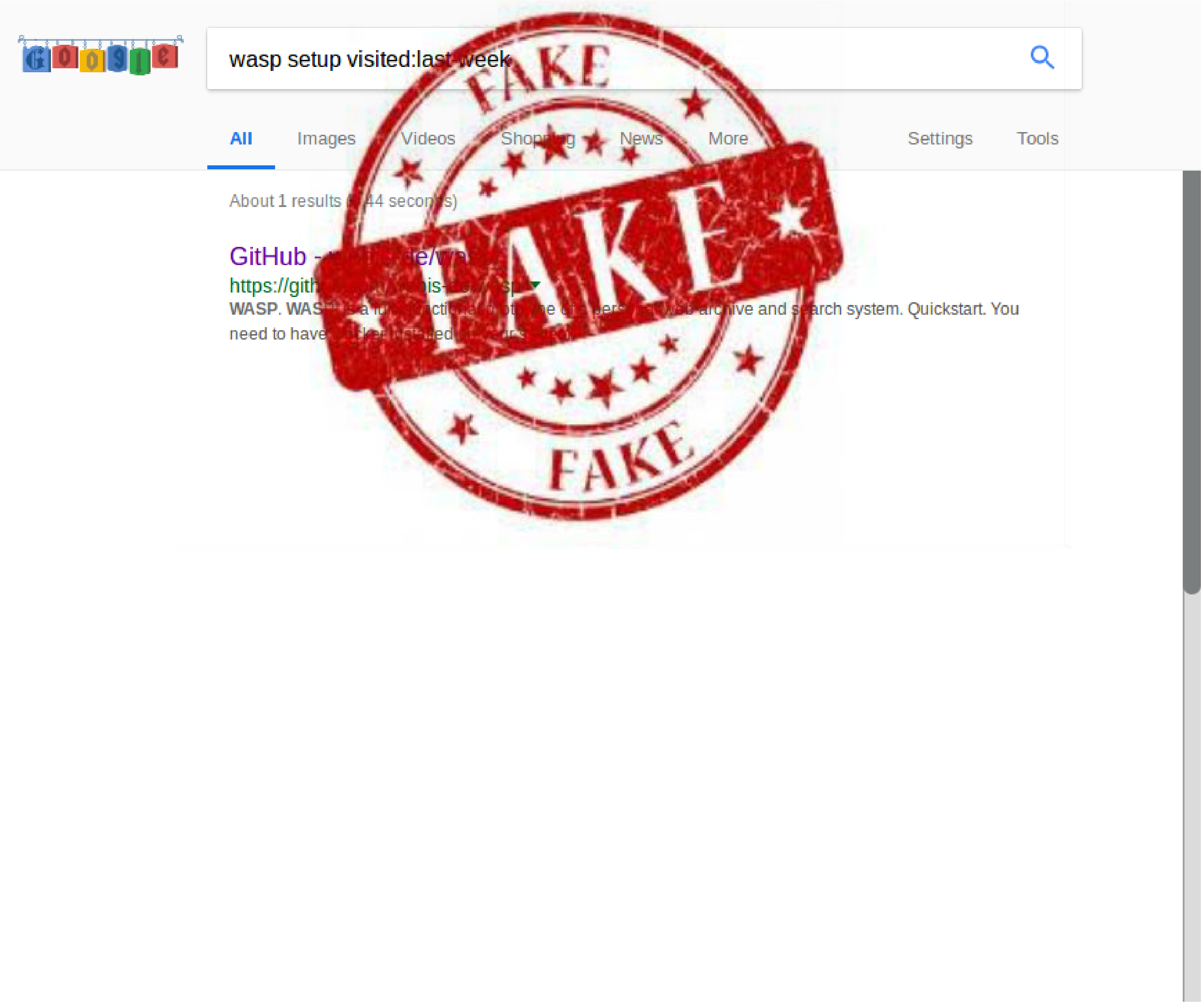


wasp setup visited:last-week

About 1 results (0,44 seconds)

GitHub - webis-de/wasp
https://github.com/webis-de/wasp ▾
**WASP**. **WASP**, is a fully functional prototype of a personal web archive and search system. Quickstart. You need to have Docker installed on your system.

# The Personal Search Engine: Motivation



wasp setup visited:last-week

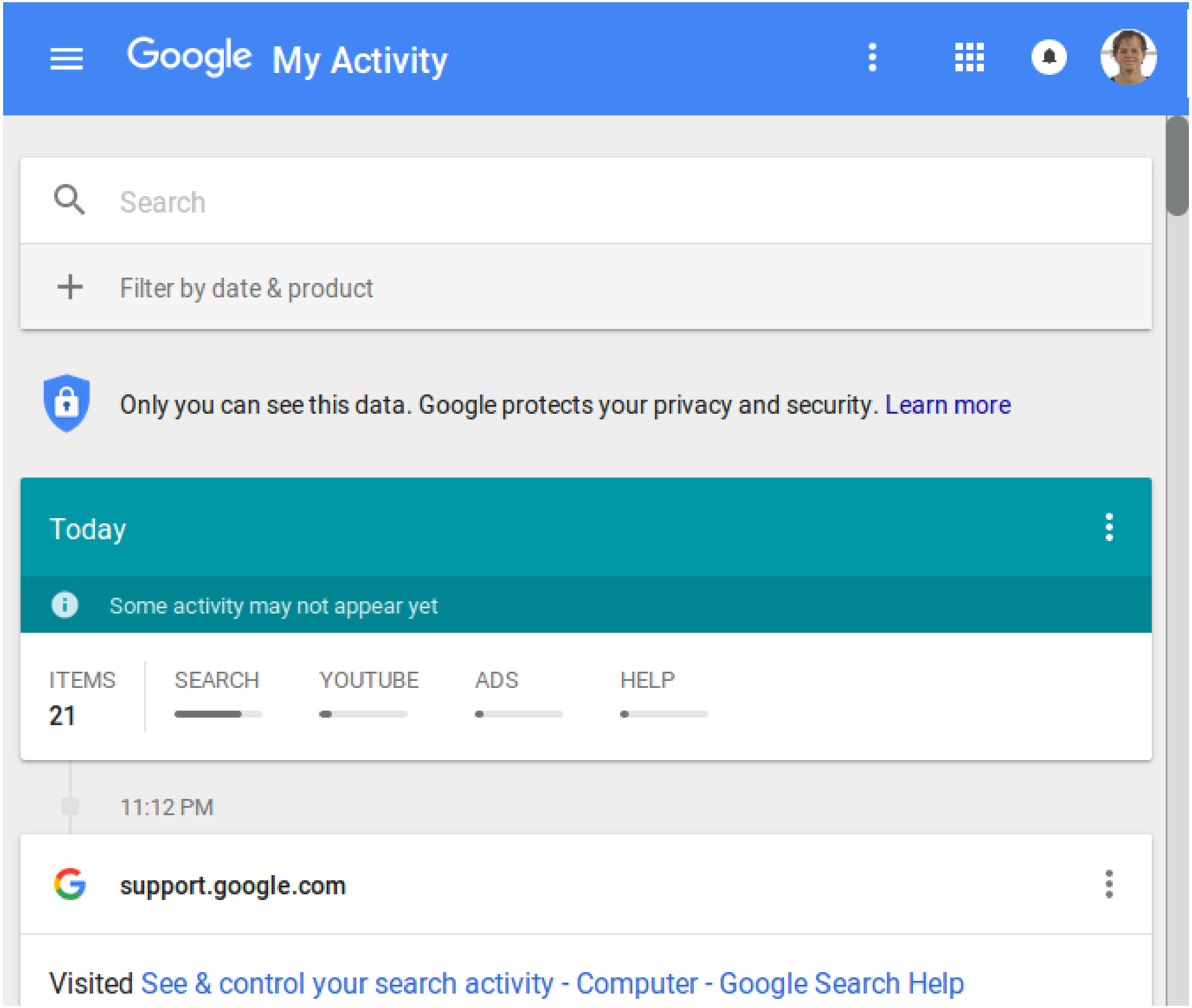
About 1 results (0.44 seconds)

GitHub - webis-de/wasp

https://github.com/webis-de/wasp

WASP. WASP is a fully functional prototype of a personal web archive and search system. Quickstart. You need to have docker installed [illegible]

# The Personal Search Engine: Motivation



Only you can see this data. Google protects your privacy and security. Learn more

Today

Some activity may not appear yet

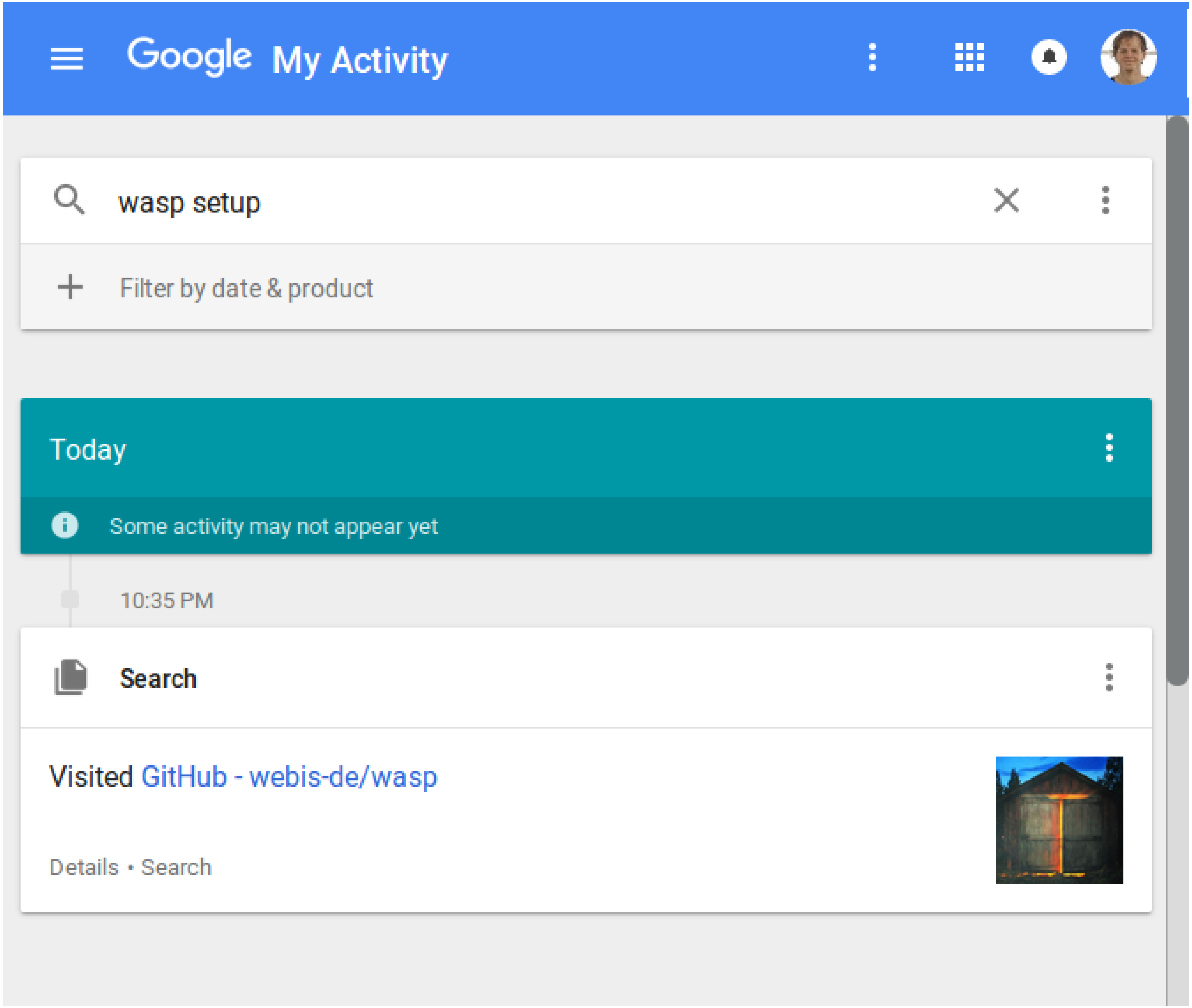
| ITEMS | SEARCH | YOUTUBE | ADS | HELP |
|---|---|---|---|---|
| 21 | | | | |

11:12 PM

support.google.com

Visited See & control your search activity - Computer - Google Search Help

# The Personal Search Engine: Motivation



wasp setup

Today

Some activity may not appear yet

10:35 PM

Search

Visited GitHub - webis-de/wasp

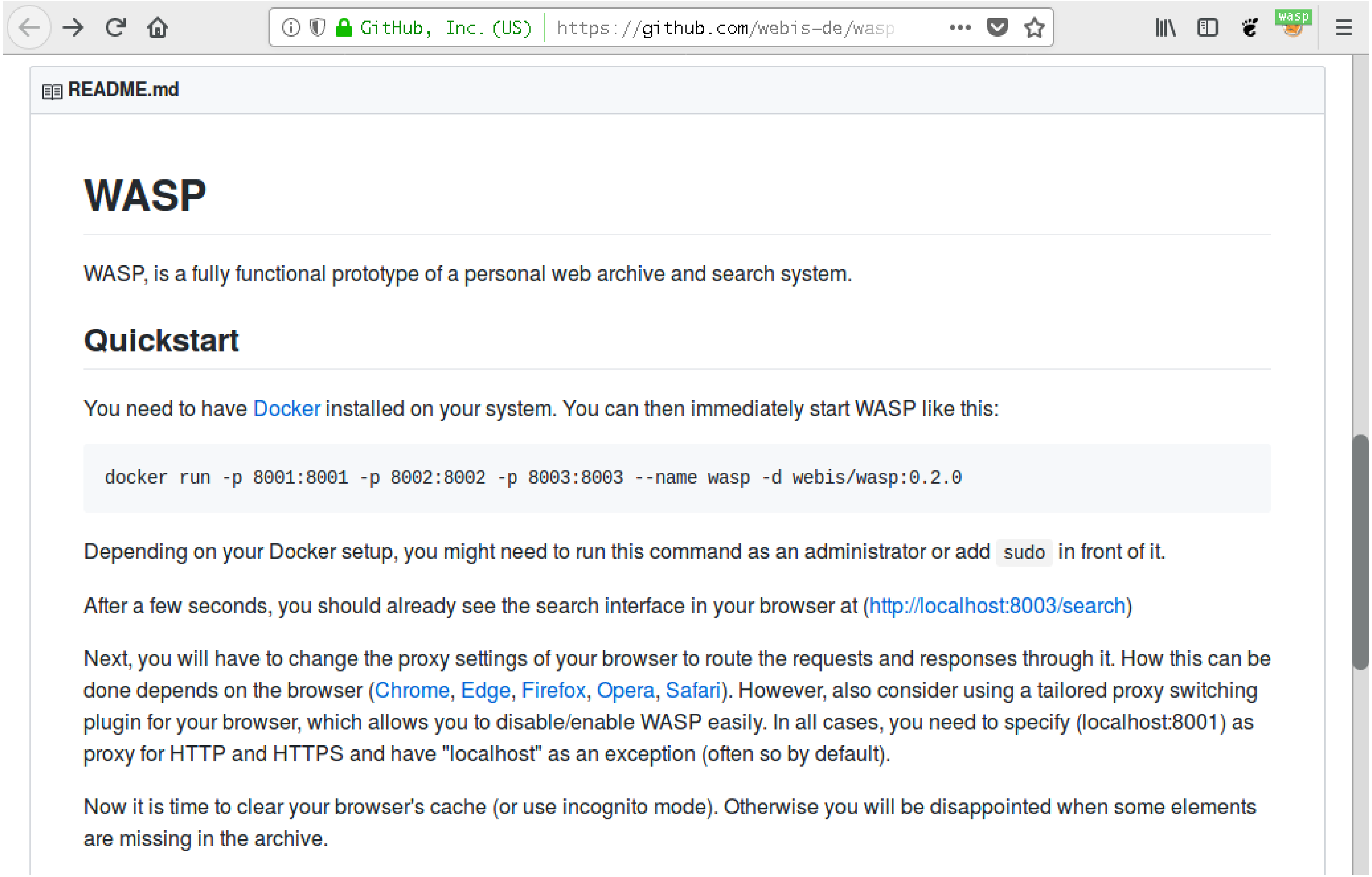
# The Personal Search Engine: Motivation

# The Personal Search Engine: Inspiration

# The Personal Search Engine: Inspiration

Personal search engine!

# WASP



# WASP

WASP, is a fully functional prototype of a personal web archive and search system.

## Quickstart

You need to have Docker installed on your system. You can then immediately start WASP like this:

```
docker run -p 8001:8001 -p 8002:8002 -p 8003:8003 --name wasp -d webis/wasp:0.2.0
```

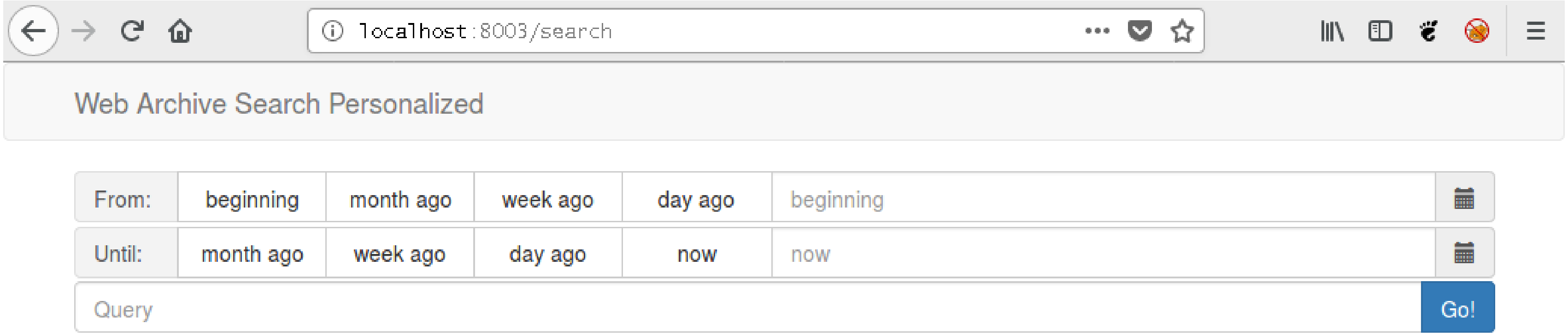
Depending on your Docker setup, you might need to run this command as an administrator or add `sudo` in front of it.

After a few seconds, you should already see the search interface in your browser at (http://localhost:8003/search)

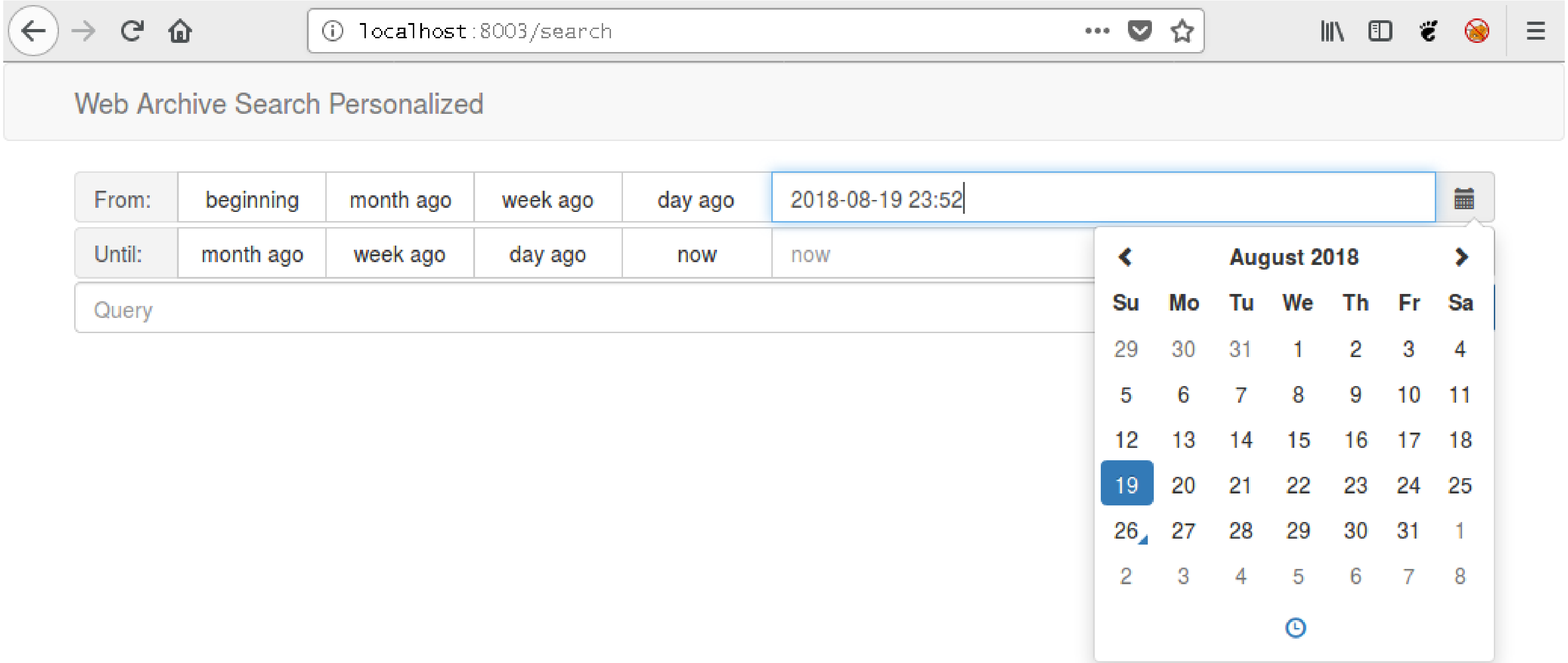
Next, you will have to change the proxy settings of your browser to route the requests and responses through it. How this can be done depends on the browser (Chrome, Edge, Firefox, Opera, Safari). However, also consider using a tailored proxy switching plugin for your browser, which allows you to disable/enable WASP easily. In all cases, you need to specify (localhost:8001) as proxy for HTTP and HTTPS and have "localhost" as an exception (often so by default).

Now it is time to clear your browser's cache (or use incognito mode). Otherwise you will be disappointed when some elements are missing in the archive.

# WASP


localhost:8003/search
Web Archive Search Personalized
From:
beginning
month ago
week ago
day ago
beginning
Until:
month ago
week ago
day ago
now
now
Query
Go!

# WASP



localhost:8003/search

Web Archive Search Personalized

From: 2018-08-19 23:52

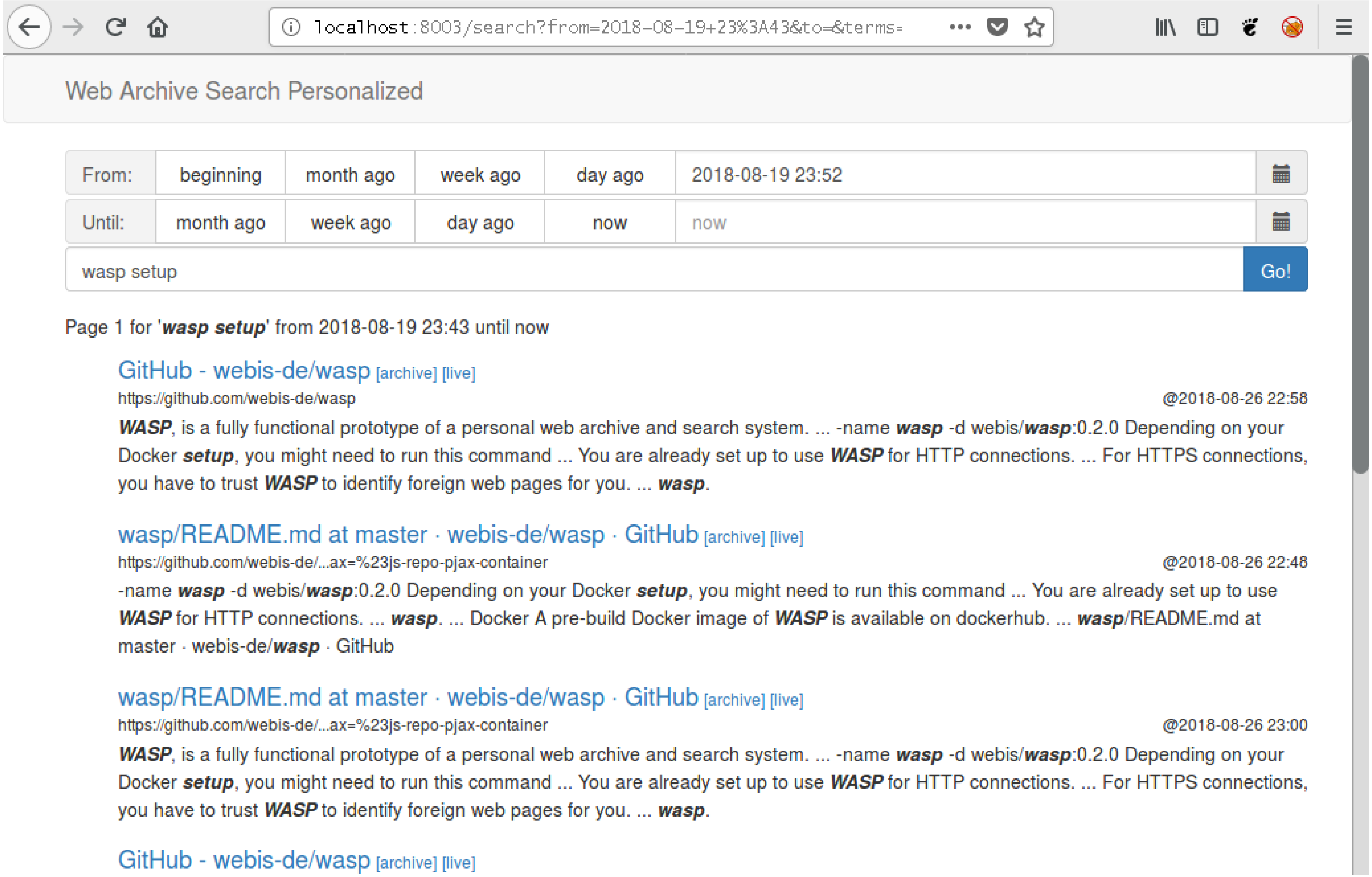
Until: now

Query

August 2018

| Su | Mo | Tu | We | Th | Fr | Sa |
|---|---|---|---|---|---|---|
| 29 | 30 | 31 | 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| 26 | 27 | 28 | 29 | 30 | 31 | 1 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |

# WASP



localhost:8003/search?from=2018-08-19+23%3A43&to=&terms=

Web Archive Search Personalized

From: 2018-08-19 23:52

Until: now

wasp setup

Page 1 for '***wasp setup***' from 2018-08-19 23:43 until now

GitHub - webis-de/wasp [archive] [live]
https://github.com/webis-de/wasp @2018-08-26 22:58

***WASP***, is a fully functional prototype of a personal web archive and search system. ... -name ***wasp*** -d webis/***wasp***:0.2.0 Depending on your Docker ***setup***, you might need to run this command ... You are already set up to use ***WASP*** for HTTP connections. ... For HTTPS connections, you have to trust ***WASP*** to identify foreign web pages for you. ... ***wasp***.

wasp/README.md at master · webis-de/wasp · GitHub [archive] [live]
https://github.com/webis-de/...ax=%23js-repo-pjax-container @2018-08-26 22:48

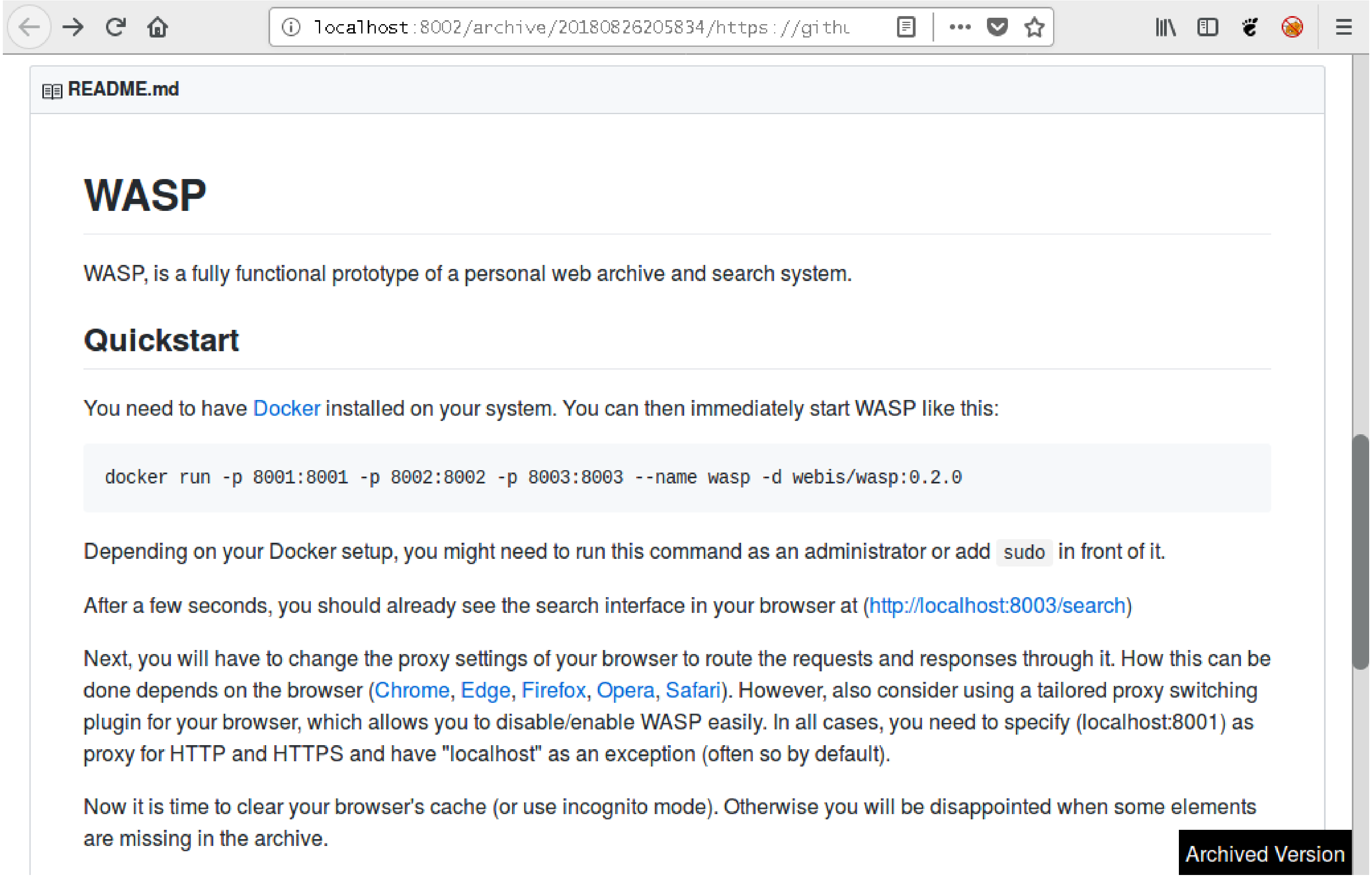
-name ***wasp*** -d webis/***wasp***:0.2.0 Depending on your Docker ***setup***, you might need to run this command ... You are already set up to use ***WASP*** for HTTP connections. ... ***wasp***. ... Docker A pre-build Docker image of ***WASP*** is available on dockerhub. ... ***wasp***/README.md at master · webis-de/***wasp*** · GitHub

wasp/README.md at master · webis-de/wasp · GitHub [archive] [live]
https://github.com/webis-de/...ax=%23js-repo-pjax-container @2018-08-26 23:00

***WASP***, is a fully functional prototype of a personal web archive and search system. ... -name ***wasp*** -d webis/***wasp***:0.2.0 Depending on your Docker ***setup***, you might need to run this command ... You are already set up to use ***WASP*** for HTTP connections. ... For HTTPS connections, you have to trust ***WASP*** to identify foreign web pages for you. ... ***wasp***.

GitHub - webis-de/wasp [archive] [live]

# WASP



# WASP

WASP, is a fully functional prototype of a personal web archive and search system.

## Quickstart

You need to have Docker installed on your system. You can then immediately start WASP like this:

```
docker run -p 8001:8001 -p 8002:8002 -p 8003:8003 --name wasp -d webis/wasp:0.2.0
```

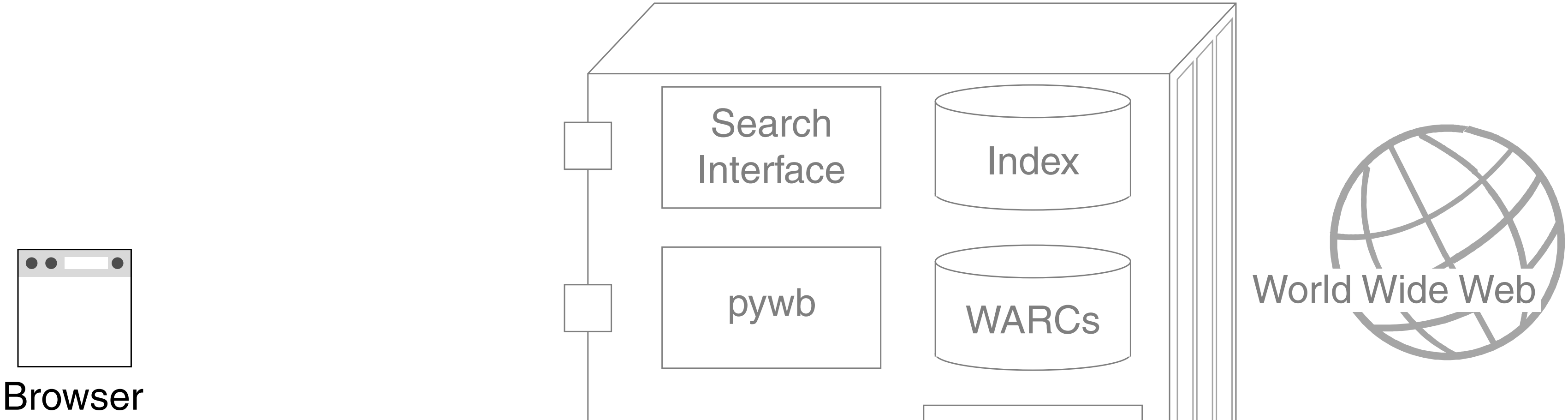
Depending on your Docker setup, you might need to run this command as an administrator or add `sudo` in front of it.

After a few seconds, you should already see the search interface in your browser at (http://localhost:8003/search)

Next, you will have to change the proxy settings of your browser to route the requests and responses through it. How this can be done depends on the browser (Chrome, Edge, Firefox, Opera, Safari). However, also consider using a tailored proxy switching plugin for your browser, which allows you to disable/enable WASP easily. In all cases, you need to specify (localhost:8001) as proxy for HTTP and HTTPS and have "localhost" as an exception (often so by default).

Now it is time to clear your browser's cache (or use incognito mode). Otherwise you will be disappointed when some elements are missing in the archive.

# WASP



Browser

Search Interface

Index

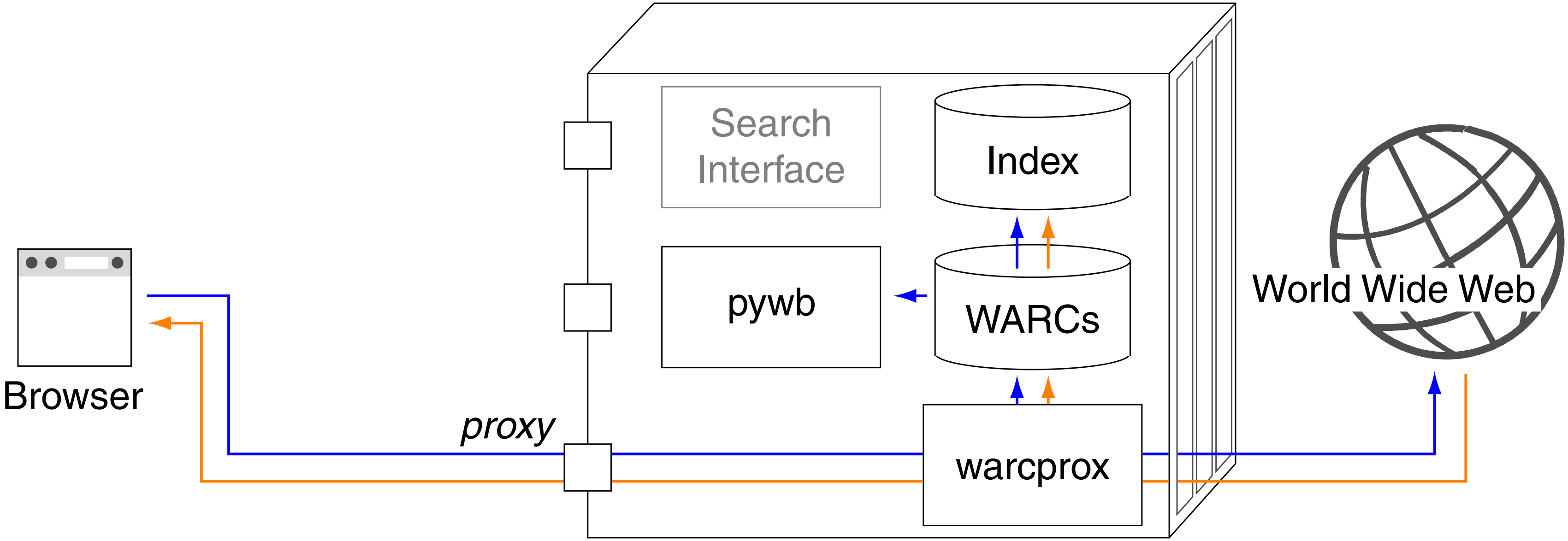
pywb

WARCs

warcprox

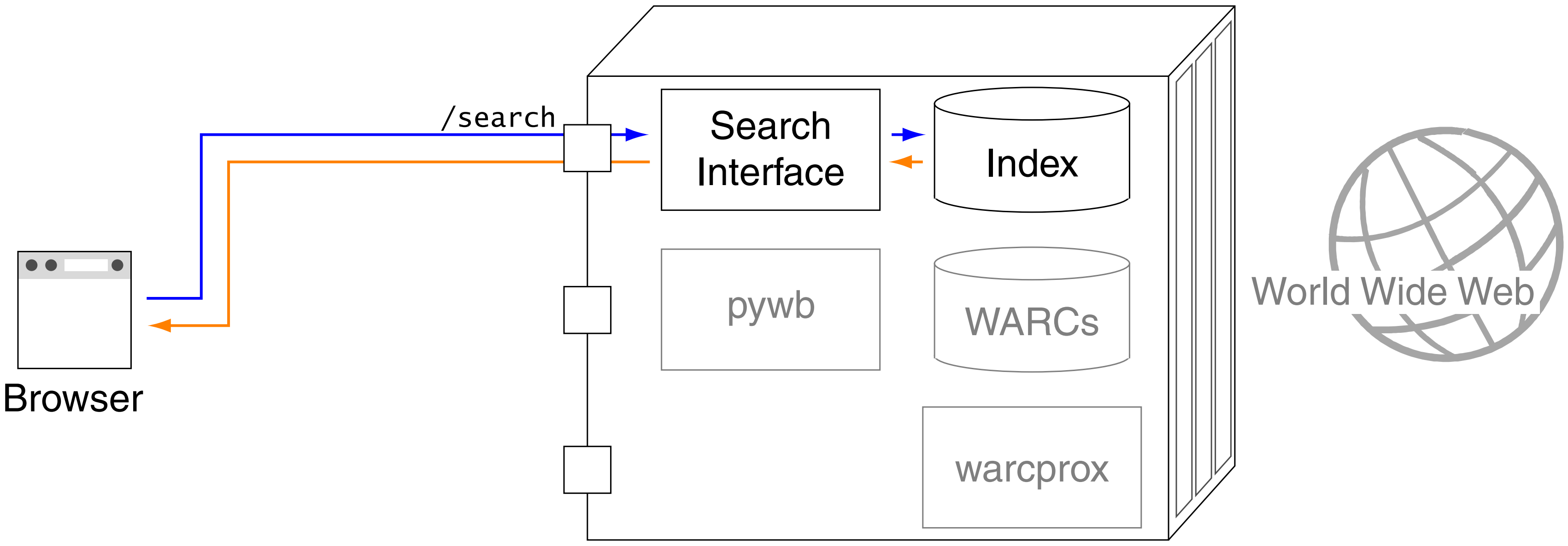
World Wide Web

# WASP



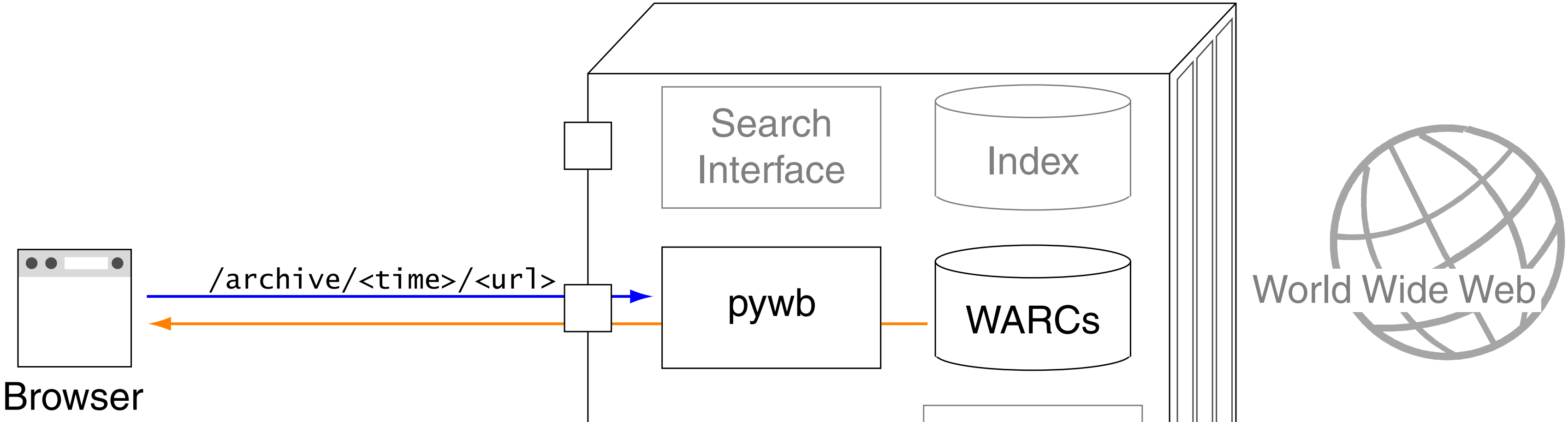
- All requests (→) and responses (←) while browsing are stored and indexed

# WASP



- ❑ All requests (→) and responses (←) while browsing are stored and indexed
- ❑ Page on localhost allows to search. Result page links to archive where...

# WASP

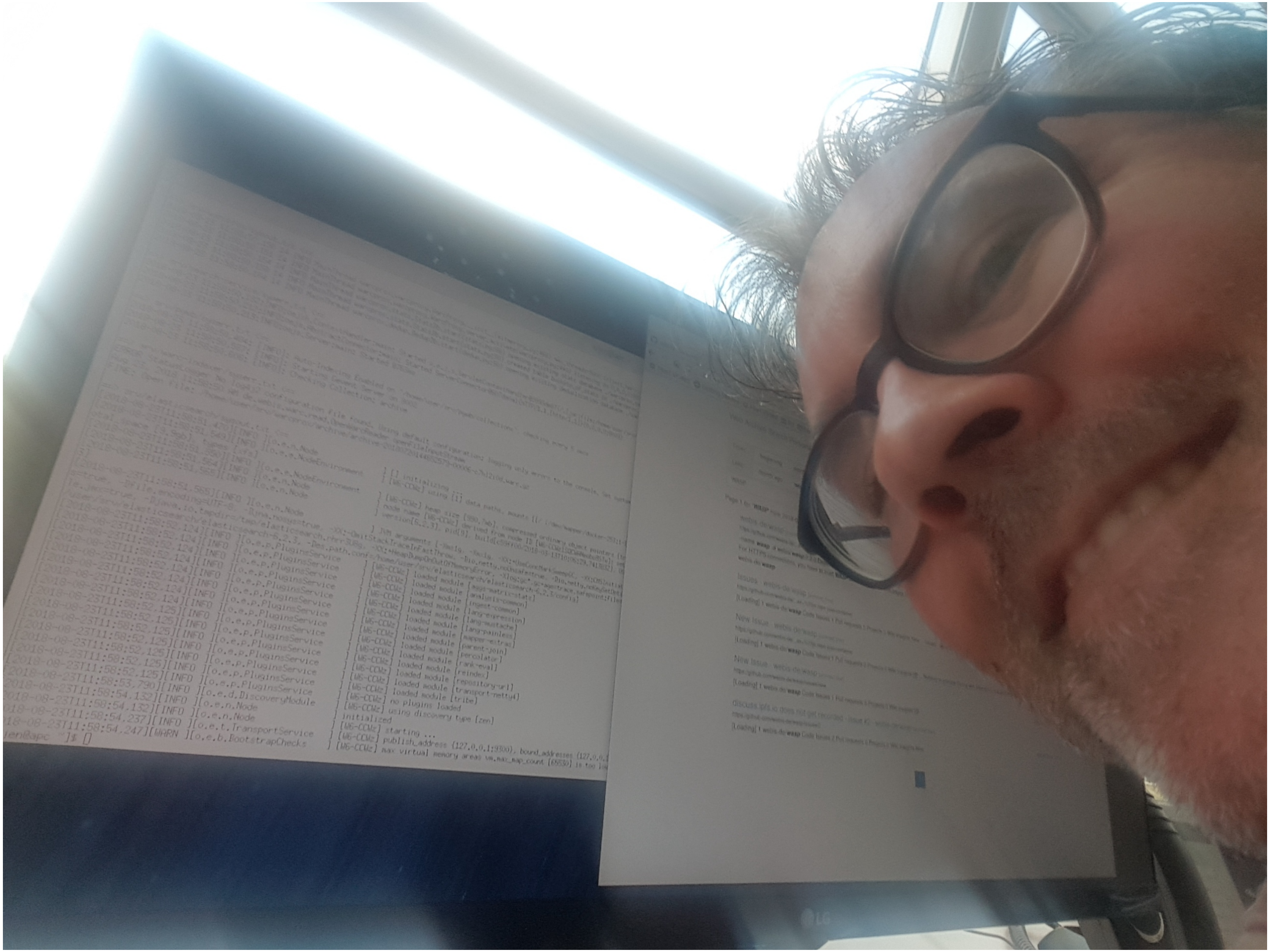


- All requests (→) and responses (←) while browsing are stored and indexed
- Page on localhost allows to search. Result page links to archive where...
- Visited pages are reproduced for the corresponding time

# WASP

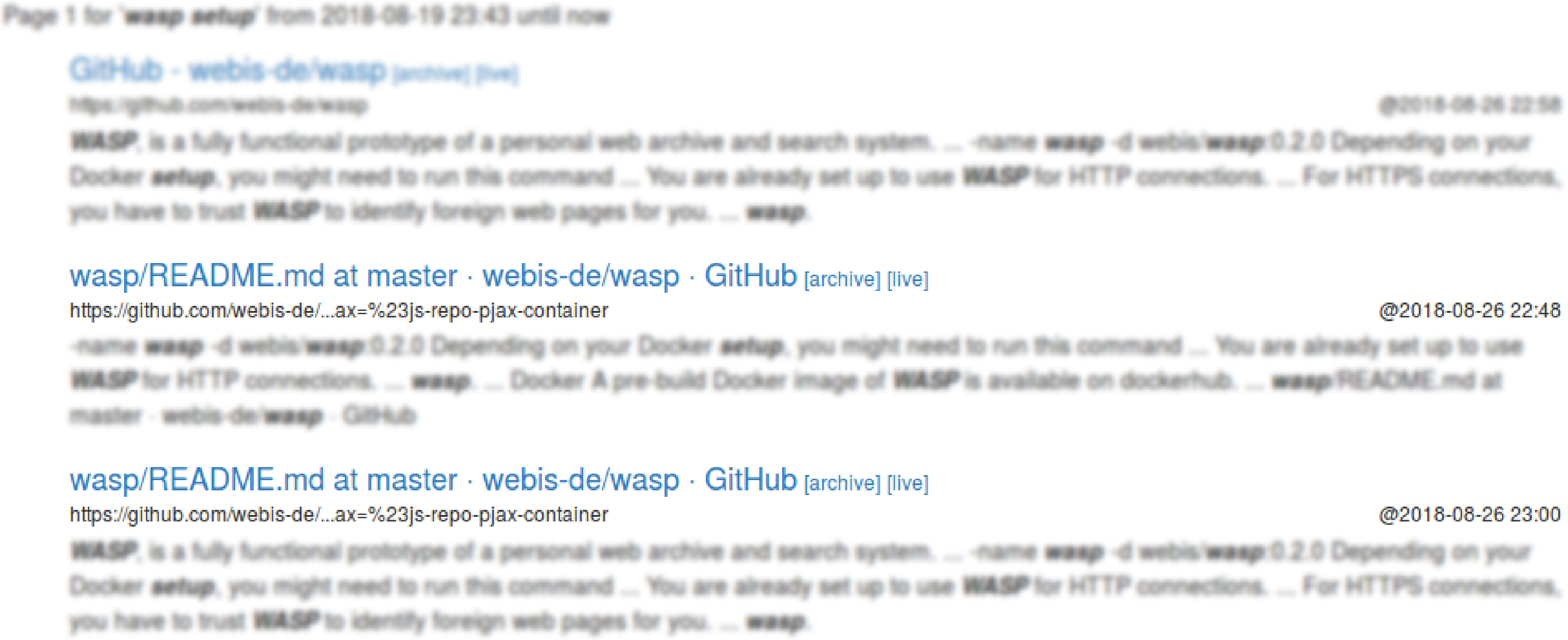
Personal
search engine!

# WASP



```
[2018-08-23T11:58:51.565][INFO ][o.e.n.Node               ] JVM arguments [-Xms1g, -Xmx1g, -XX:+UseConcMarkSweepGC, -XX:CMSInitiat
se=true, -Dfile.encoding=UTF-8, -Djna.nosys=true, -XX:-OmitStackTraceInFastThrow, -Dio.netty.noUnsafe=true, -Dio.netty.noKeySetOpt
le.jmx=true, -Djava.io.tmpdir=/tmp/elasticsearch.rhrr3U8g, -XX:+HeapDumpOnOutOfMemoryError, -Xlog:gc*,gc+age=trace,safepoint:file
/user/srv/elasticsearch/elasticsearch-6.2.3, -Des.path.conf=/home/user/srv/elasticsearch/elasticsearch-6.2.3/config]
[2018-08-23T11:58:52.124][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [aggs-matrix-stats]
[2018-08-23T11:58:52.124][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [analysis-common]
[2018-08-23T11:58:52.124][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [ingest-common]
[2018-08-23T11:58:52.124][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [lang-expression]
[2018-08-23T11:58:52.124][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [lang-mustache]
[2018-08-23T11:58:52.124][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [lang-painless]
[2018-08-23T11:58:52.125][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [mapper-extras]
[2018-08-23T11:58:52.125][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [parent-join]
[2018-08-23T11:58:52.125][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [percolator]
[2018-08-23T11:58:52.125][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [rank-eval]
[2018-08-23T11:58:52.125][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [reindex]
[2018-08-23T11:58:52.125][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [repository-url]
[2018-08-23T11:58:52.125][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [transport-netty4]
[2018-08-23T11:58:52.125][INFO ][o.e.p.PluginsService     ] [W6-CCWz] loaded module [tribe]
[2018-08-23T11:58:52.125][INFO ][o.e.p.PluginsService     ] [W6-CCWz] no plugins loaded
[2018-08-23T11:58:53.790][INFO ][o.e.d.DiscoveryModule    ] [W6-CCWz] using discovery type [zen]
[2018-08-23T11:58:54.132][INFO ][o.e.n.Node               ] initialized
[2018-08-23T11:58:54.132][INFO ][o.e.n.Node               ] [W6-CCWz] starting ...
[2018-08-23T11:58:54.237][INFO ][o.e.t.TransportService   ] [W6-CCWz] publish_address {127.0.0.1:9300}, bound_addresses {127.0.0.1
[2018-08-23T11:58:54.247][WARN ][o.e.b.BootstrapChecks    ] [W6-CCWz] max virtual memory areas vm.max_map_count [65530] is too low
jen@apc ~]$ []
```

# Insight 1: Not Indexing Near-duplicates



Page 1 for '**wasp setup**' from 2018-08-19 23:43 until now

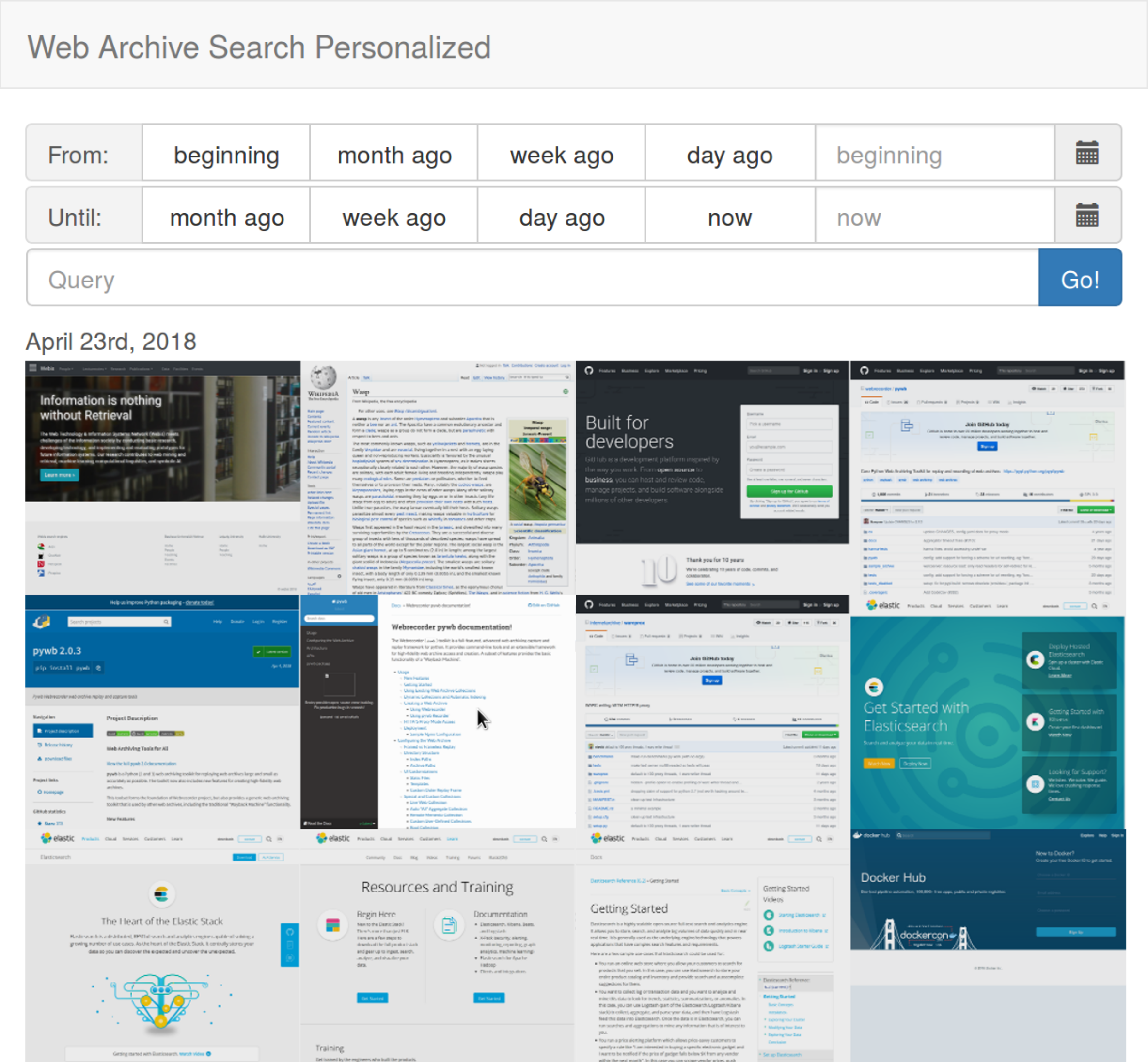
GitHub - webis-de/wasp [archive] [live]
https://github.com/webis-de/wasp @2018-08-26 22:58
**WASP**, is a fully functional prototype of a personal web archive and search system. ... -name **wasp** -d webis/**wasp**:0.2.0 Depending on your Docker **setup**, you might need to run this command ... You are already set up to use **WASP** for HTTP connections. ... For HTTPS connections, you have to trust **WASP** to identify foreign web pages for you. ... **wasp**.

wasp/README.md at master · webis-de/wasp · GitHub [archive] [live]
https://github.com/webis-de/...ax=%23js-repo-pjax-container @2018-08-26 22:48
-name **wasp** -d webis/**wasp**:0.2.0 Depending on your Docker **setup**, you might need to run this command ... You are already set up to use **WASP** for HTTP connections. ... **wasp**. ... Docker A pre-build Docker image of **WASP** is available on dockerhub. ... **wasp**/README.md at master · webis-de/**wasp** · GitHub

wasp/README.md at master · webis-de/wasp · GitHub [archive] [live]
https://github.com/webis-de/...ax=%23js-repo-pjax-container @2018-08-26 23:00
**WASP**, is a fully functional prototype of a personal web archive and search system. ... -name **wasp** -d webis/**wasp**:0.2.0 Depending on your Docker **setup**, you might need to run this command ... You are already set up to use **WASP** for HTTP connections. ... For HTTPS connections, you have to trust **WASP** to identify foreign web pages for you. ... **wasp**.

- What changes warrant a re-archiving?

# Insight 2: Browsable and Deletable History



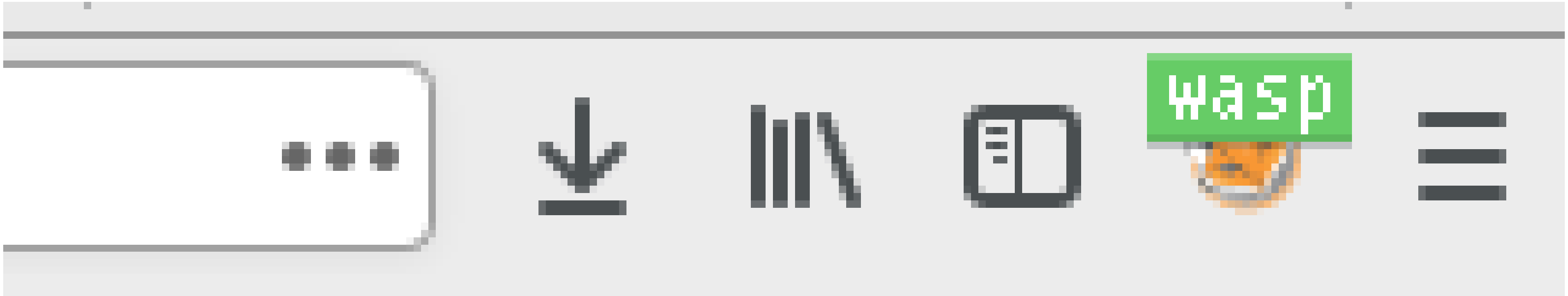
Web Archive Search Personalized

April 23rd, 2018

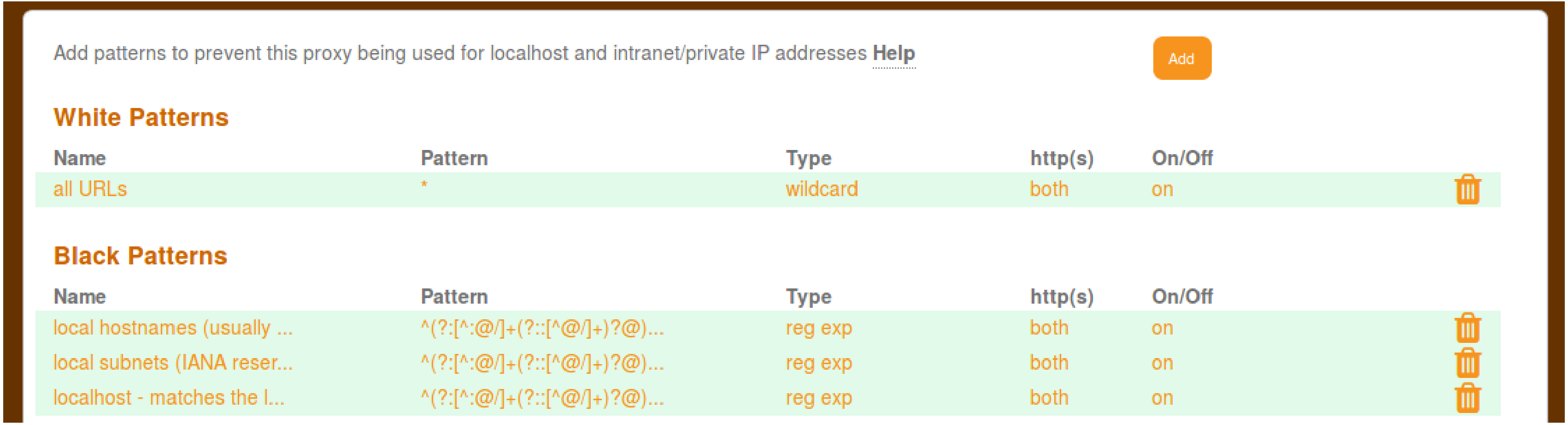
Information is nothing without Retrieval
Wasp
Built for developers
pywb 2.0.3
Webrecorder pywb documentation!
Get Started with Elasticsearch
The Heart of the Elastic Stack
Resources and Training
Getting Started
Docker Hub

# Insight 3: Easy (De-)activation of Archiving

Easy activation, deactivation, and status-check



Patterns



Add patterns to prevent this proxy being used for localhost and intranet/private IP addresses **Help**

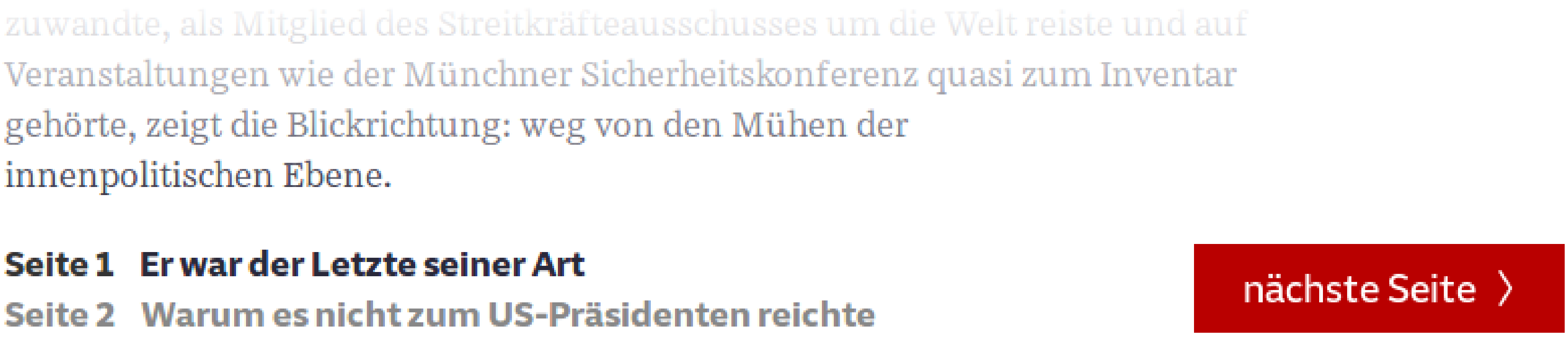
### White Patterns

| Name | Pattern | Type | http(s) | On/Off |
|---|---|---|---|---|
| all URLs | * | wildcard | both | on |

### Black Patterns

| Name | Pattern | Type | http(s) | On/Off |
|---|---|---|---|---|
| local hostnames (usually ... | ^(?:[^:@/]+(?::[^@/]+)?@)... | reg exp | both | on |
| local subnets (IANA reser... | ^(?:[^:@/]+(?::[^@/]+)?@)... | reg exp | both | on |
| localhost - matches the l... | ^(?:[^:@/]+(?::[^@/]+)?@)... | reg exp | both | on |

# Insight 4: Combined Indexing of Sub-pages



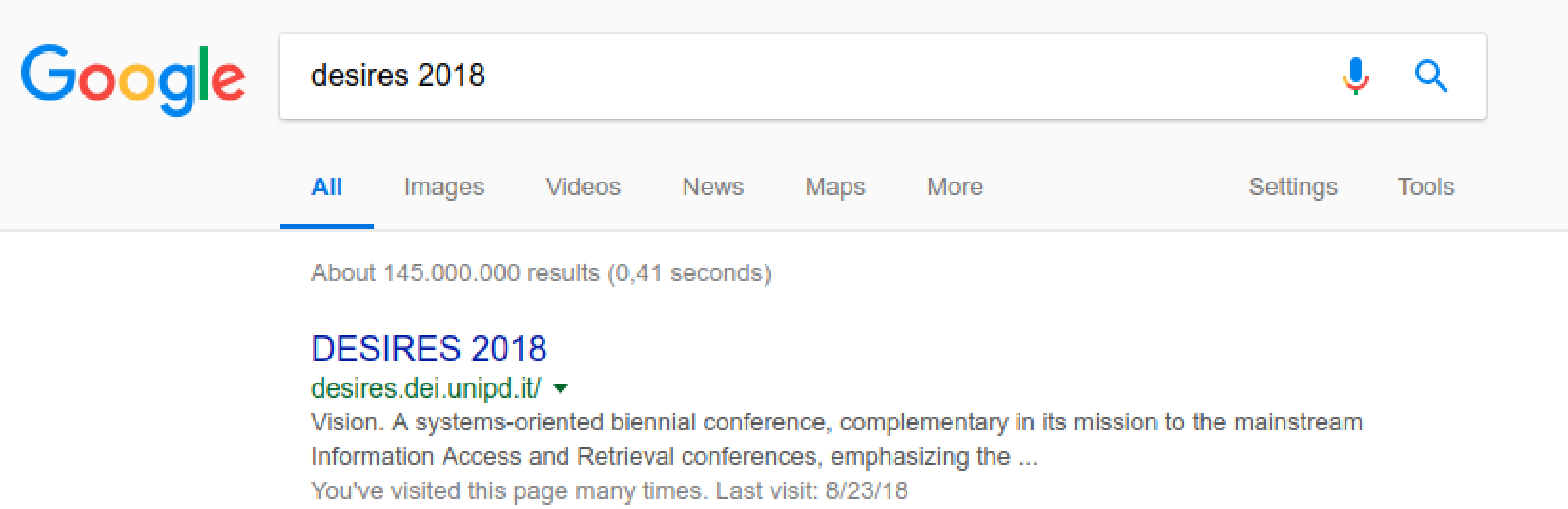
zuwandte, als Mitglied des Streitkräfteausschusses um die Welt reiste und auf Veranstaltungen wie der Münchner Sicherheitskonferenz quasi zum Inventar gehörte, zeigt die Blickrichtung: weg von den Mühen der innenpolitischen Ebene.

**Seite 1 Er war der Letzte seiner Art**
Seite 2 Warum es nicht zum US-Präsidenten reichte

❑ Should visited sub-pages of a single article be indexed as one?

❑ If so, to which sub-page should be linked in the result list?

# Insight 5: Personalized Search



desires 2018

About 145.000.000 results (0,41 seconds)

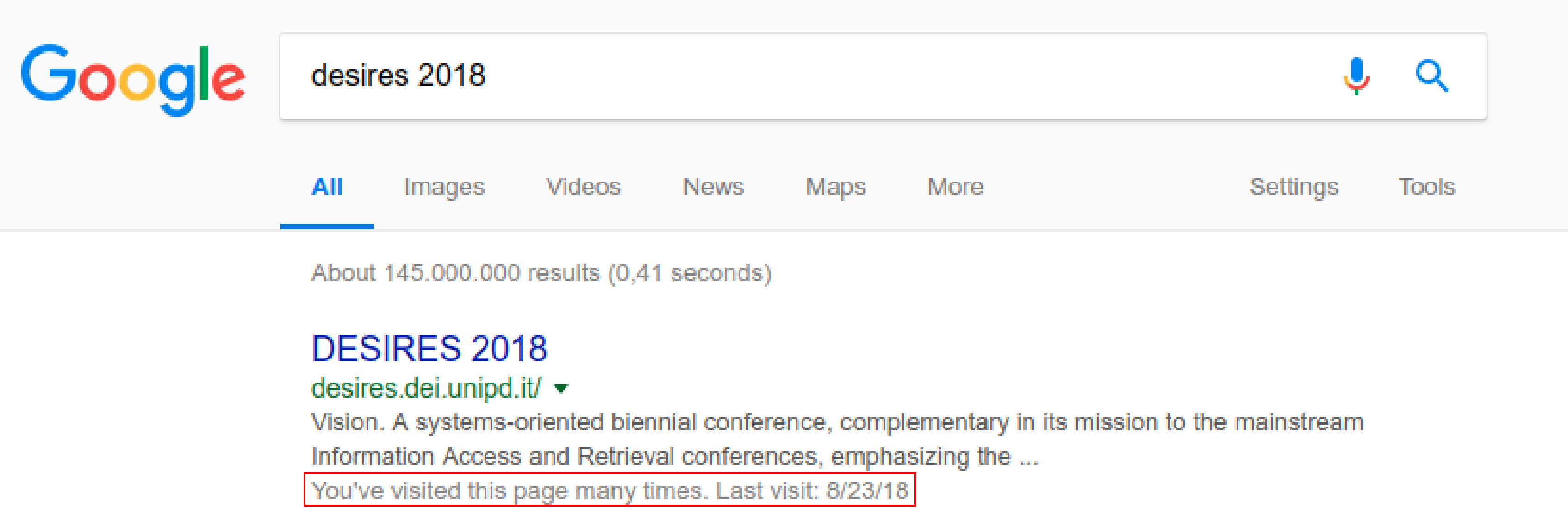
DESIRES 2018

desires.dei.unipd.it/

Vision. A systems-oriented biennial conference, complementary in its mission to the mainstream Information Access and Retrieval conferences, emphasizing the ...

You've visited this page many times. Last visit: 8/23/18

# Insight 5: Personalized Search



About 145.000.000 results (0,41 seconds)

DESIRES 2018

desires.dei.unipd.it/

Vision. A systems-oriented biennial conference, complementary in its mission to the mainstream Information Access and Retrieval conferences, emphasizing the ...

You've visited this page many times. Last visit: 8/23/18

You've visited this page many times. Last visit: 8/23/18

# WASP: Web Archiving and Search Personalized

Insights overview

- Not Indexing Near-duplicates
- Browsable and Deletable History
- Easy (De-)activation of Archiving
- Combined Indexing of Sub-pages
- Personalized Search

Code and Instructions on Github

`github.com/webis-de/wasp`

Thank you for your attention!