

WASP: Web Archiving and Search Personalized

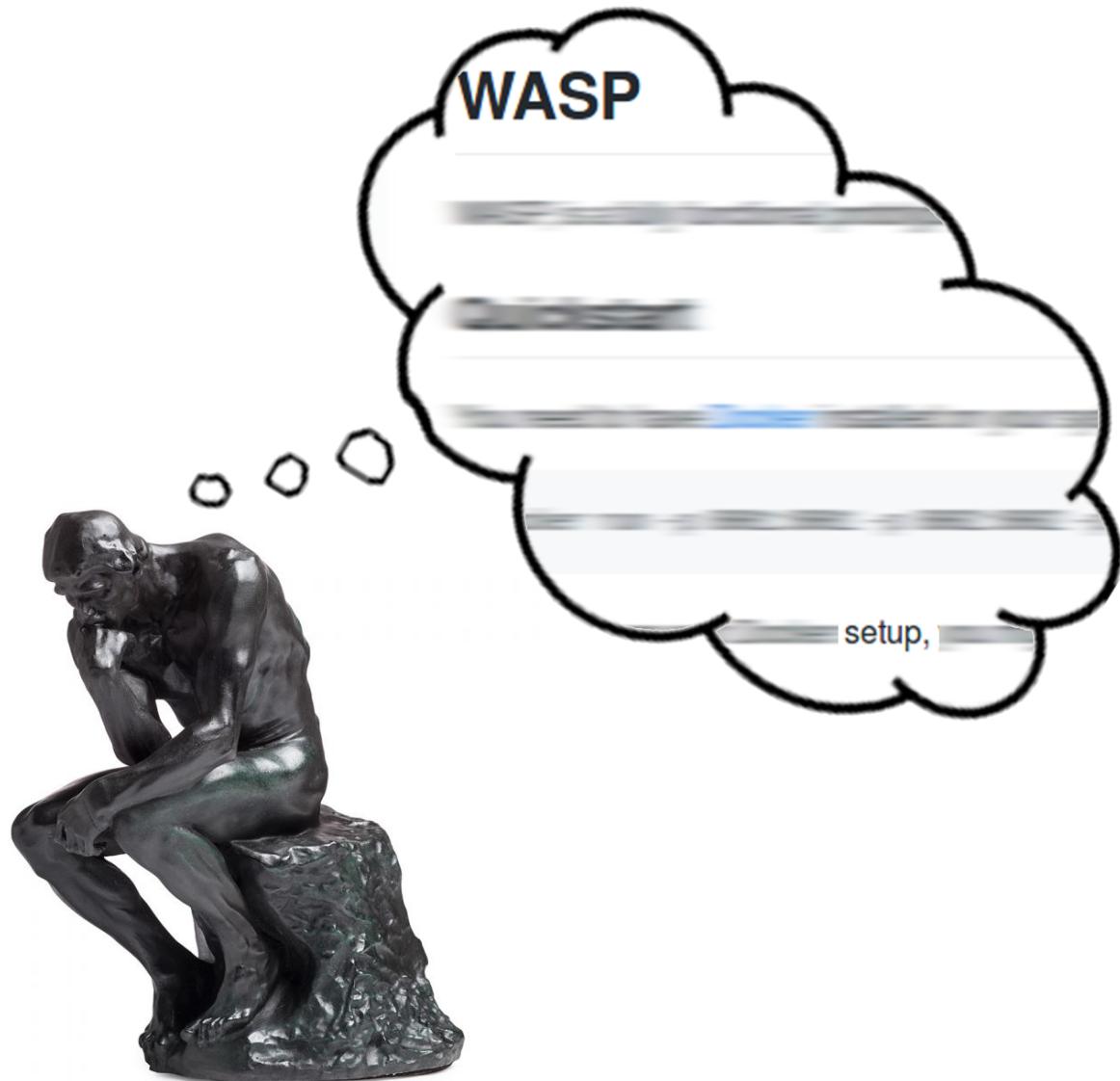
Johannes Kiesel, Arjen P. de Vries, Matthias Hagen, Benno Stein and Martin Potthast
@KieselJohannes, @arjenpdevries, @matthias_hagen, @bennostein, @martinpottbast

DESires, August 29th 2018

The Personal Search Engine: Motivation



The Personal Search Engine: Motivation



The Personal Search Engine: Motivation

🕒 History ▾ ✕

🔍 Search history

View ▾

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

The Personal Search Engine: Motivation

🕒 History ▾

✖

🔍 Search history

View ▾

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

🕒 History ▾

✖

🔍 wasp

✖

View ▾

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

The Personal Search Engine: Motivation

🕒 History ▾ X

🔍 Search history View▼

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

🕒 History ▾ X

🔍 wasp X View▼

- G webis wasp – Google Search
- G wasp – Google Search
- G wasp webis – Google-Suche
- 🐦 Johannes Kiesel auf Twitter:...
- ⌚ webis-de/wasp
- G Google – My Activity
- 💻 WASP: Web Archiving and Sear...
- G Google Advanced Search
- G 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 - ..
- G amber wasp – Google Search
- PIX Wasp Insect Sting - Free ima...

🕒 History ▾ X

🔍 wasp setup X View▼

The Personal Search Engine: Motivation

A screenshot of a search engine results page. At the top, the Google logo is visible next to a search bar containing the query "wasp setup". Below the search bar are navigation links for "All", "Images", "Videos", "Shopping", "News", and "More", with "All" being underlined. To the right are "Settings" and "Tools" links. The search results indicate "About 2.950.000 results (0,44 seconds)". The first result is a link titled "Getting Started Training - Wasp Barcode" with the URL <https://www.waspbarcode.com/training/getting-started-training>. A snippet below the link states: "Available for MobileAsset, WaspTime and InventoryControl, training covers basics including installation, database setup, and overview of key features and icons." Below this, there is a section titled "Videos" featuring three video thumbnails. The first video is titled "WASP Basic Install and basic features Demo" by "waspAnalytics" (YouTube - Jul 1, 2010). The second video is titled "Unboxing and set up of the Wasp Delta 20X40" by "Simone | FNTSMN" (YouTube - Feb 21, 2016). The third video is titled "How to set up the WHY Trap for Wasps, Hornets & Yellowjackets" by "WhyRescue" (YouTube - Jul 15, 2017). Each video thumbnail includes its duration: 2:00, 9:01, and 1:32 respectively. A right-pointing arrow is located to the right of the third video's title.

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

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

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

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

A screenshot of a web-based search interface. At the top, there is a navigation bar with tabs for "All", "Images", "Videos", "Shopping", "News", and "More". To the right of the tabs are "Settings" and "Tools" buttons. Below the navigation bar, a search bar contains the query "wasp setup visited:last-week". A magnifying glass icon is positioned to the right of the search bar. The main content area displays the search results. It starts with a message "About 1 results (0.44 seconds)". Below this, a single result is shown: a link to "GitHub - webis-de/wasp" with the URL "https://github.com/webis-de/wasp". A dropdown arrow next to the URL indicates more information. Below the link, a brief description reads: "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



The Personal Search Engine: Motivation

The screenshot shows the Google My Activity interface. At the top, there's a blue header bar with the title "Google My Activity". Below it is a search bar labeled "Search" and a button to "Filter by date & product". A note about privacy and security is displayed, stating "Only you can see this data. Google protects your privacy and security." with a link to learn more. The main area is titled "Today" and shows a message "Some activity may not appear yet". It lists 21 items, with categories for SEARCH, YOUTUBE, ADS, and HELP. A timestamp "11:12 PM" is shown above a list item. The first item in the list is a visit to "support.google.com".

≡ Google My Activity

Search

+ Filter by date & product

Only you can see this data. Google protects your privacy and security. [Learn more](#)

Today

Some activity may not appear yet

ITEMS 21 | SEARCH | YOUTUBE | ADS | HELP

11:12 PM

G support.google.com

Visited [See & control your search activity - Computer - Google Search Help](#)

The Personal Search Engine: Motivation

≡ Google My Activity ⋮ ⋮ ⚡ 📺

🔍 wasp setup ✖ ⋮

+ Filter by date & product

Today ⋮

ⓘ Some activity may not appear yet

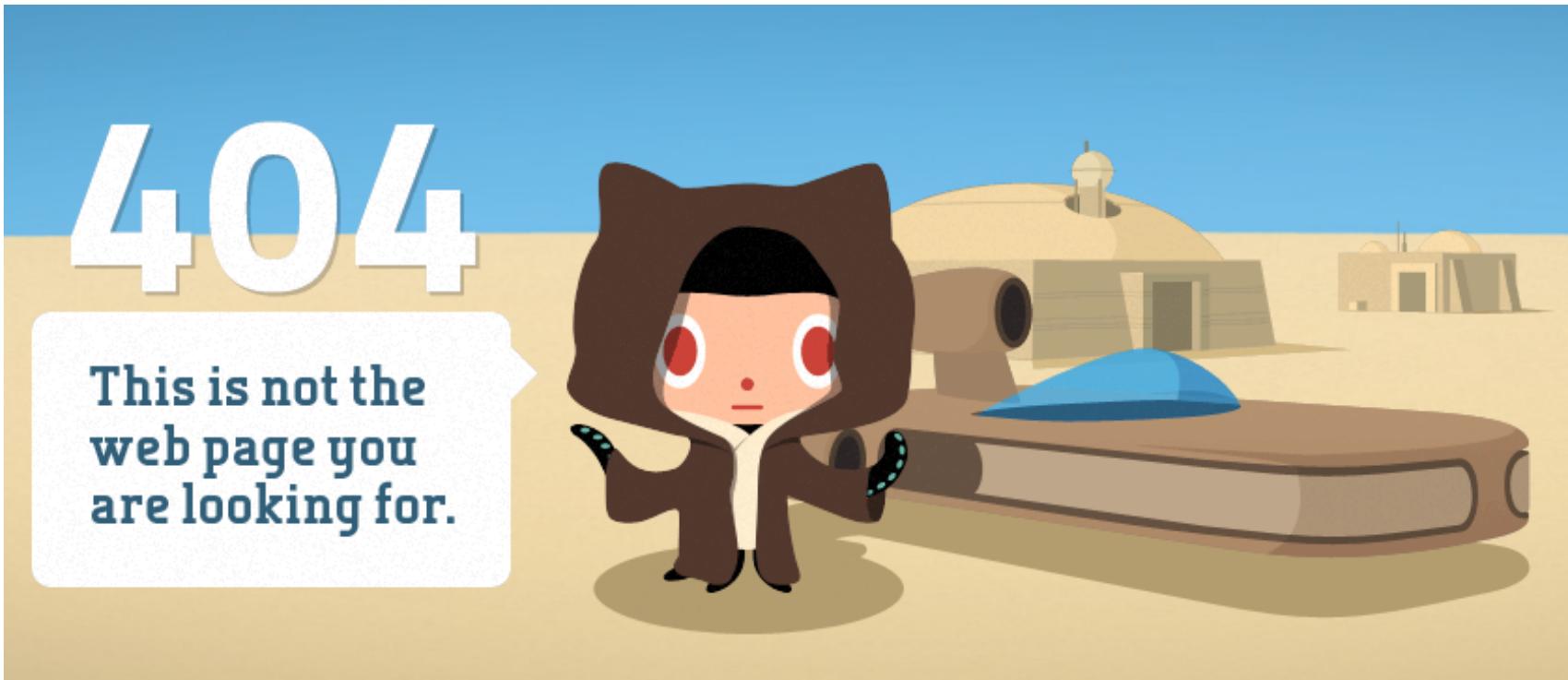
10:35 PM

📄 Search ⋮

Visited GitHub - webis-de/wasp

Details • Search 

The Personal Search Engine: Motivation



The Personal Search Engine: Inspiration



The Personal Search Engine: Inspiration



Personal
search engine!

The screenshot shows a web browser window with the URL <https://github.com/webis-de/wasp>. The page displays the contents of the `README.md` file. The title "WASP" is prominently displayed at the top. Below it, a brief description states: "WASP, is a fully functional prototype of a personal web archive and search system." A section titled "Quickstart" follows, with instructions: "You need to have Docker installed on your system. You can then immediately start WASP like this:" and a command-line example: `docker run -p 8001:8001 -p 8002:8002 -p 8003:8003 --name wasp -d webis/wasp:0.2.0`. Further instructions advise running the command as an administrator or adding `sudo`. The text continues with information about changing browser proxy settings and clearing browser cache.

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:	beginning	month ago	week ago	day ago	2018-08-19 23:52	
Until:	month ago	week ago	day ago	now	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

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

From: beginning month ago week ago day ago 2018-08-19 23:52

Until: month ago week ago day ago now now

wasp setup Go!

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]



README.md

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.

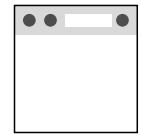
After a few seconds, you should already see the search interface in your browser at (<http://localhost:8003/search>)

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).

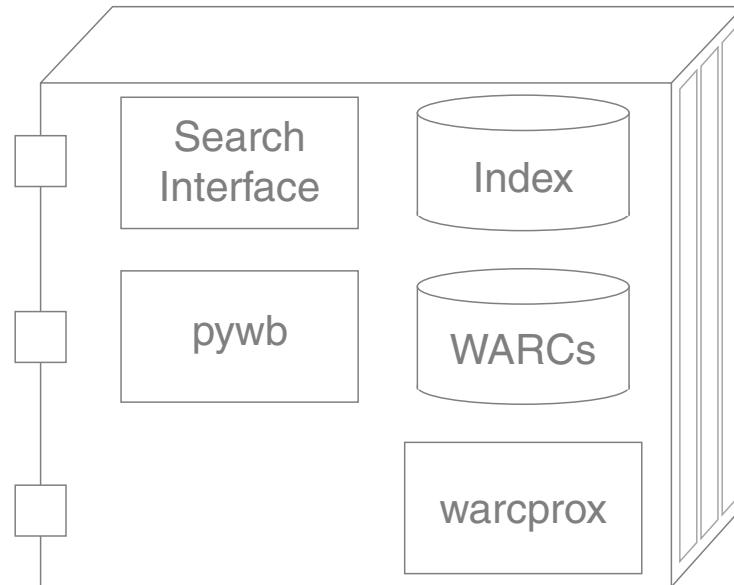
Now it is time to clear your browser's cache (or use incognito mode). Otherwise you will be disappointed when some elements are missing in the archive.

Archived Version

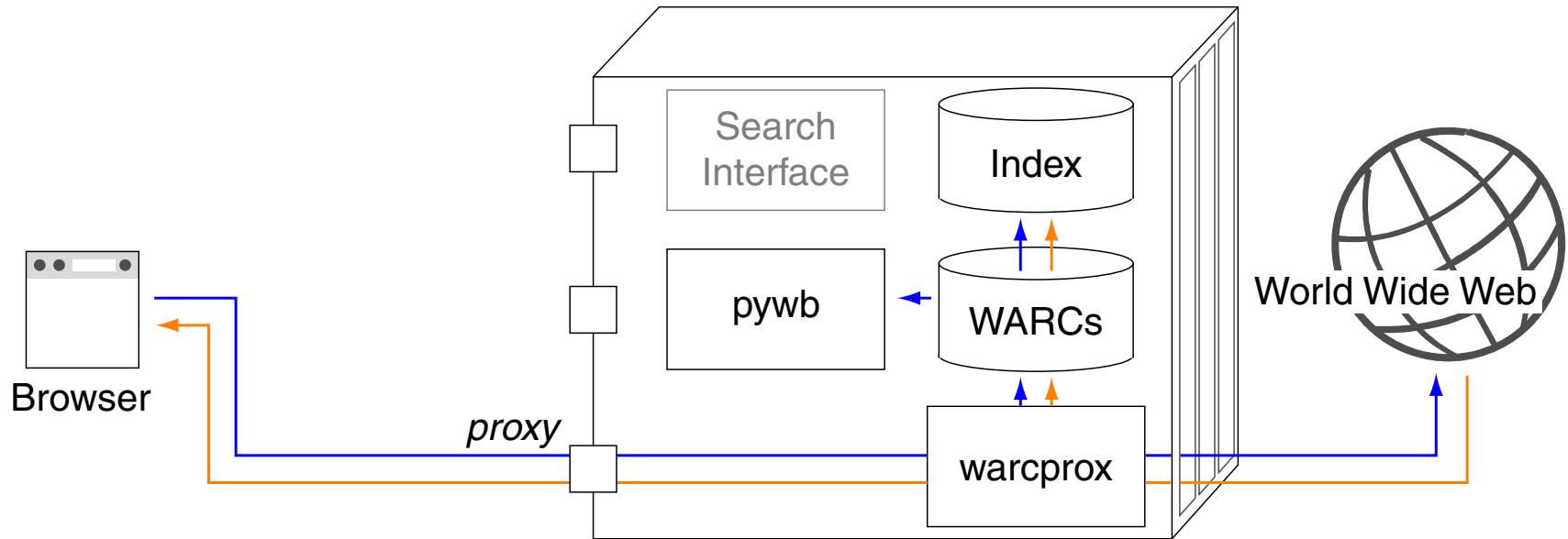
WASP



Browser

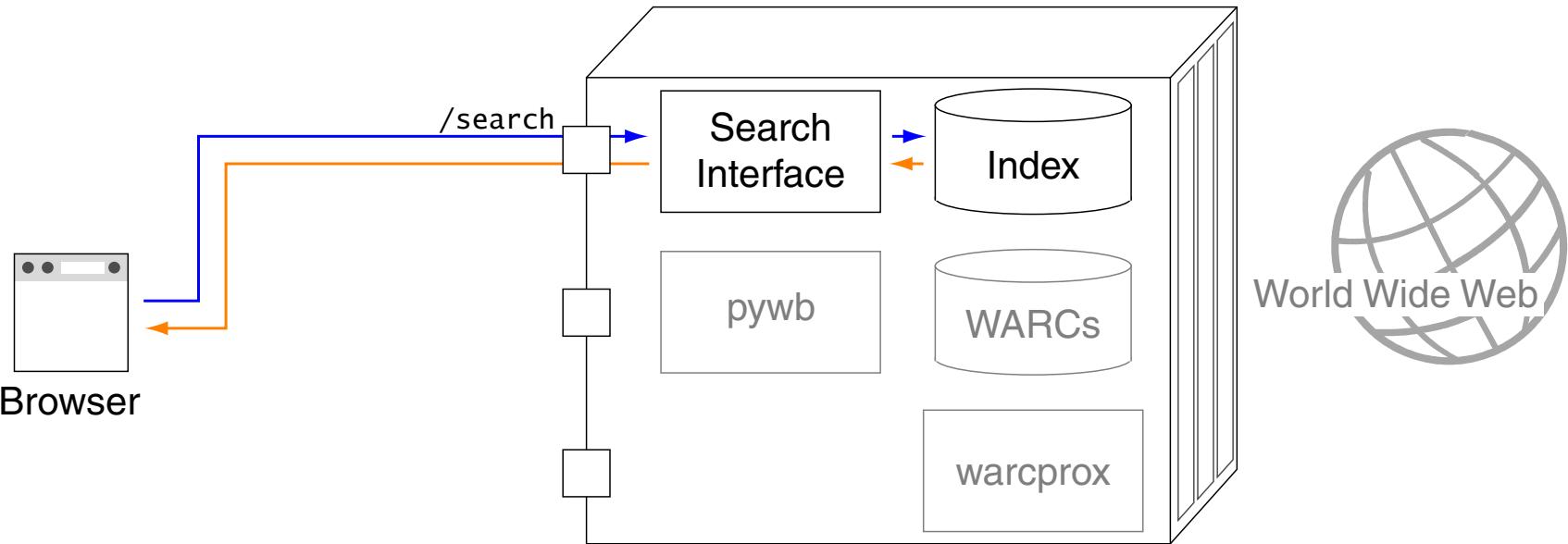


WASP



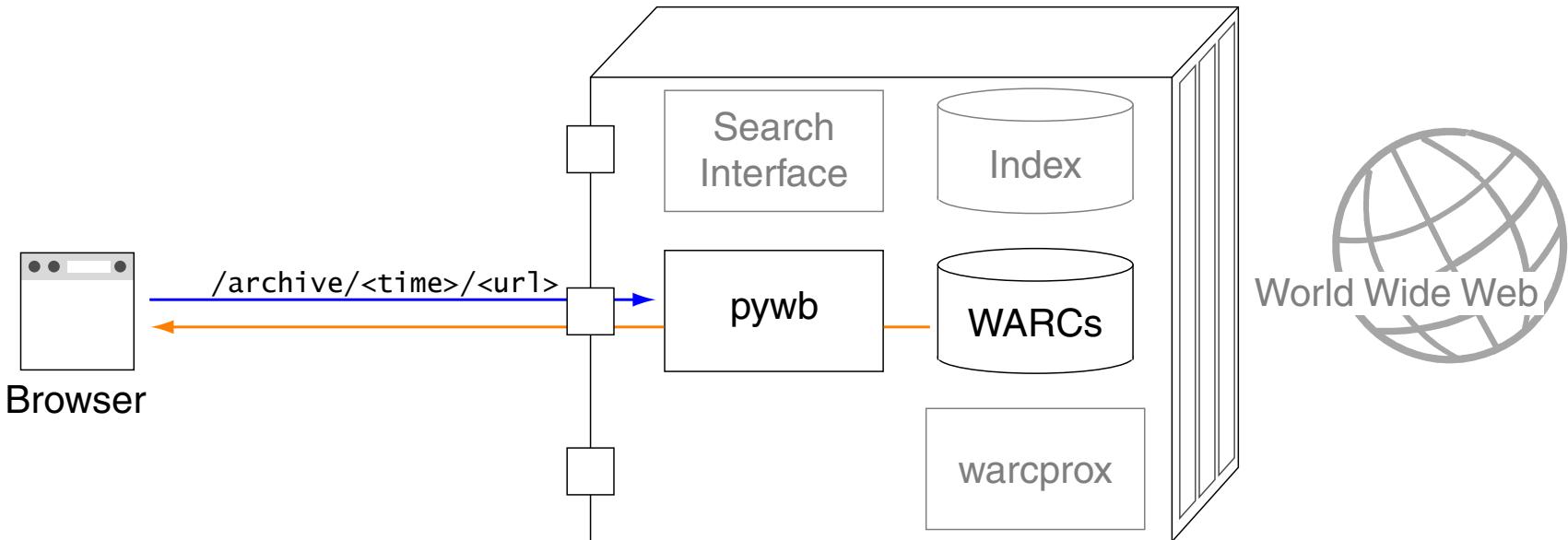
- All requests (→) and responses (←) while browsing are stored and indexed

WASP



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

WASP



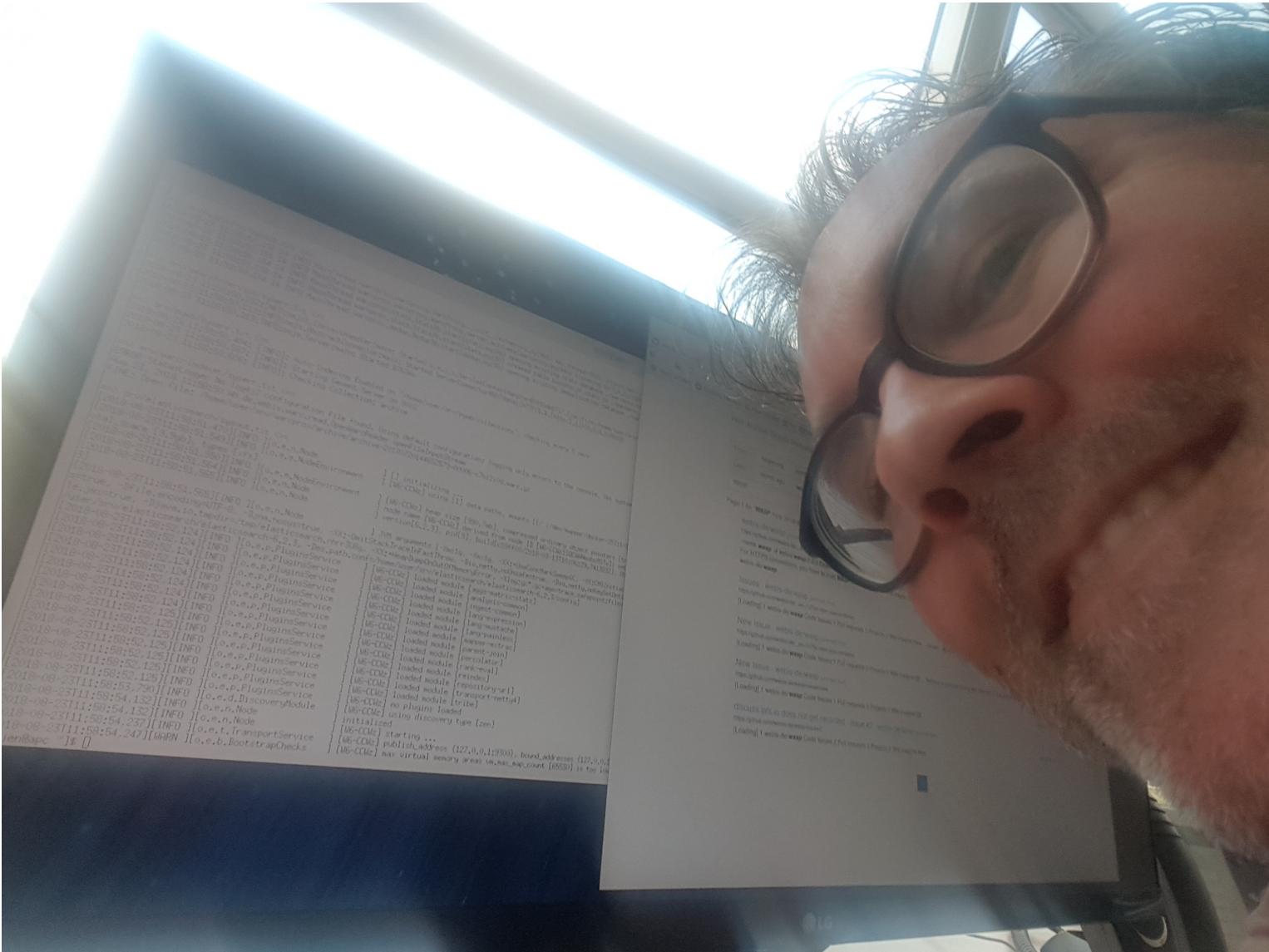
- ❑ All requests (→) and responses (←) while browsing are stored and indexed
- ❑ Page on localhost allows to search. Result page links to archive where...
- ❑ Visited pages are reproduced for the corresponding time

WASP



Personal
search engine!

WASP



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

From:	beginning	month ago	week ago	day ago	beginning	
Until:	month ago	week ago	day ago	now	now	

Query Go!

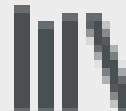
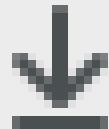
April 23rd, 2018

The grid displays nine screenshots of web archive search interfaces, likely from the Wayback Machine or similar services, showing different search results and features for specific dates in April 2018.

- Row 1:**
 - Information is nothing without Retrieval (Wayback Machine)
 - Wasp (Wayback Machine)
 - Built for developers (Wayback Machine)
- Row 2:**
 - pywb 2.0.3 (Wayback Machine)
 - pywb documentation (Wayback Machine)
 - Java 8 Update 141 (Wayback Machine)
- Row 3:**
 - Elastic Stack (Wayback Machine)
 - Resources and Training (Wayback Machine)
 - Docker Hub (Wayback Machine)

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](#)

[Add](#)

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.

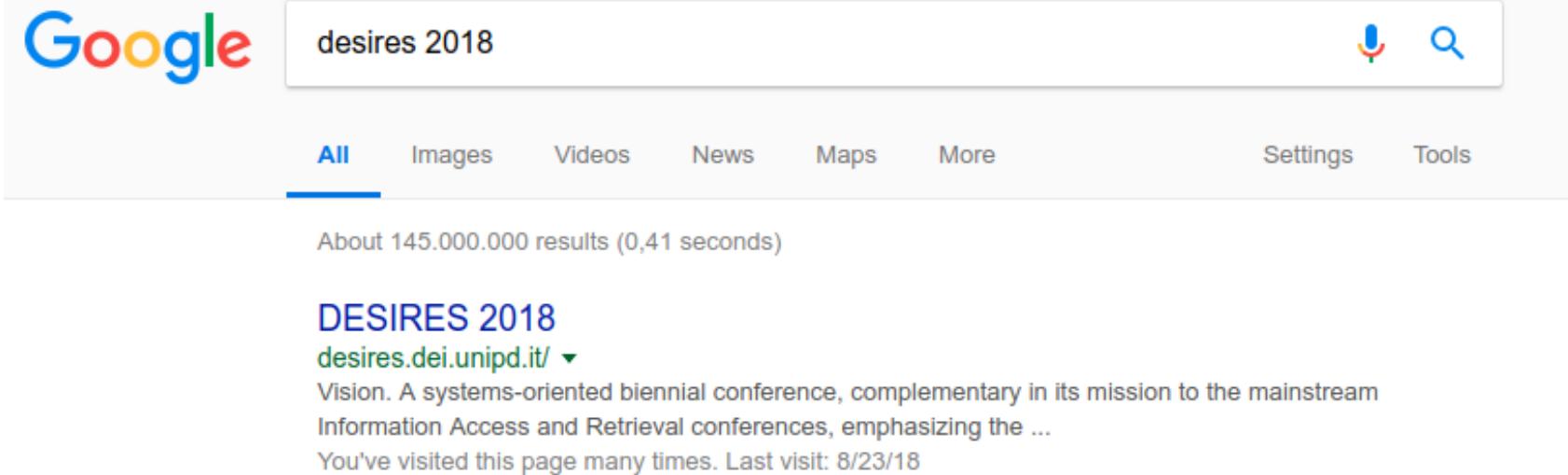
Seite 1 Er war der Letzte seiner Art

Seite 2 Warum es nicht zum US-Präsidenten reichte

nächste Seite >

- 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



A screenshot of a Google search results page. The search query "desires 2018" is entered in the search bar. The "All" tab is selected, showing approximately 145 million results found in 0.41 seconds. The top result is for "DESIRE 2018" from the website desires.dei.unipd.it/. The snippet describes it as a systems-oriented biennial conference complementary to mainstream Information Access and Retrieval conferences. A note indicates the user has visited the page many times, with the last visit on 8/23/18.

desires 2018

All Images Videos News Maps More Settings Tools

About 145.000.000 results (0,41 seconds)

DESIRE 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



desires 2018



All

Images

Videos

News

Maps

More

Settings

Tools

About 145.000.000 results (0,41 seconds)

[DESIRE 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!