State of play of the Data Acquisition Service (DAS) of the Web Intelligence Hub (WIH)

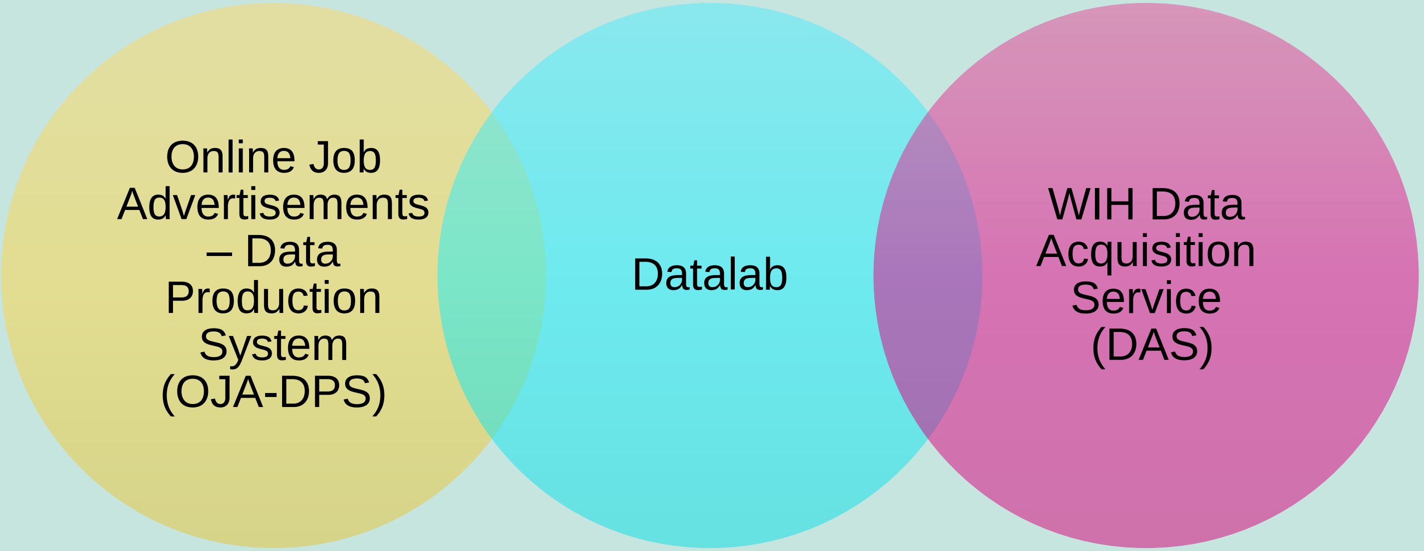
Mészáros Mátyás

Web Intelligence Network Conference - From Web to Data
Gdansk

2025/02/04

European Commission

WIH platform components



Online Job Advertisements – Data Production System (OJA-DPS)

Datalab

WIH Data Acquisition Service (DAS)

DAS development principles

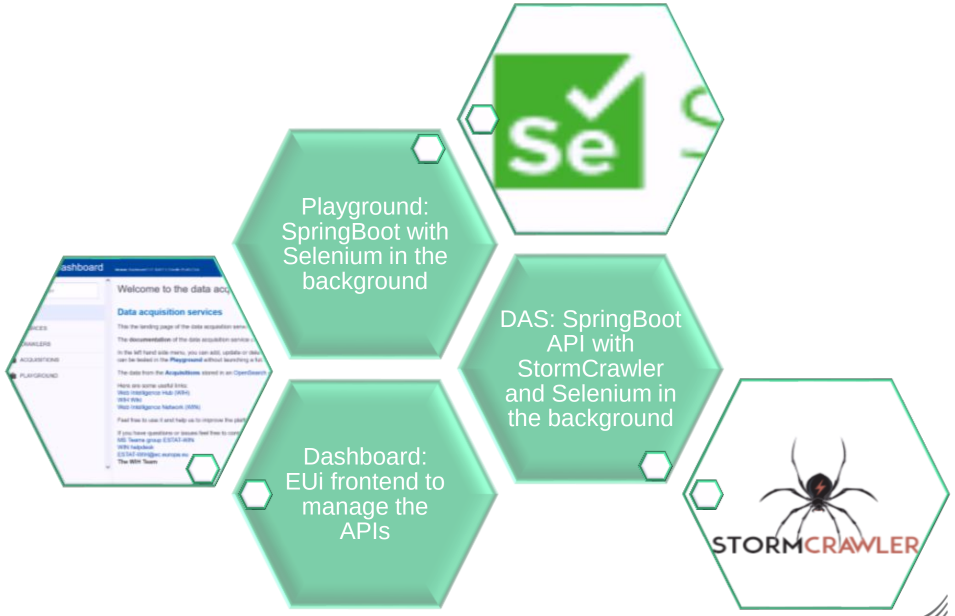
- Scalable
- Build on open-source tools
- Try to use the state of the art
- Can handle static and dynamic content
- No coding, only configuration by the user
- Universal, can be used for several use cases (OJA, MNE, price, etc.)
- Separation of use cases with possible collaboration in the same use case

The beginning of the DAS

- **Version 1** was released in 2021 Nov
  - Generic data acquisition service with API access for static and dynamic web pages
  - Using StormCrawler and Selenium
  - EUi frontend (Dashboard) with user authentication (EU Login and AWS Cognito)
  - Deployment using Infrastructure as Code (IaC)

- **Version 2** was released in 2022 Apr
  - Adding the playground for data acquisition to test filters and dynamic web pages

- **Version 3** was released in 2022 Sep
  - Additional Selenium filters based on the needs of tourism websites



Playground: SpringBoot with Selenium in the background

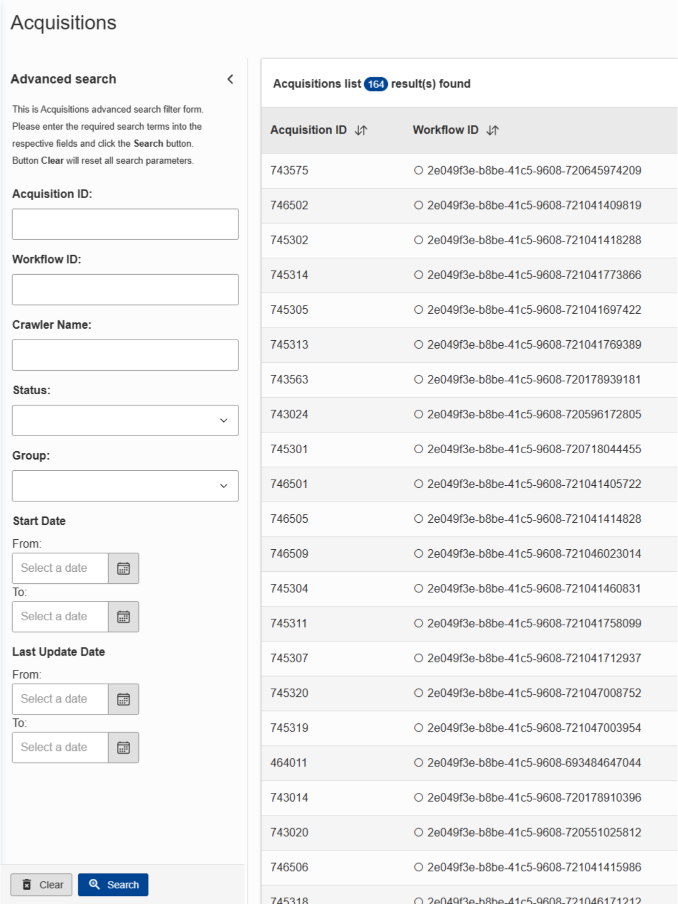
DAS: SpringBoot API with StormCrawler and Selenium in the background

Dashboard: EUi frontend to manage the APIs

First testing by the WIN

Based on the feedback

- **Version 4** was released in 2022 Dec
  - Multitenancy was introduced
  - Moving authentication from AWS Cognito to Keycloak
- **Version 5** was released in 2023 Feb
  - 1st Security testing and update
  - Adding new functionalities like advanced search, copy configuration and acquisition action history
- **Version 6** was released in 2023 Mar
  - Introduction of user roles (guest, developer and admin)
  - Possibility to use sitemap discovery



Acquisitions

Advanced search

This is Acquisitions advanced search filter form. Please enter the required search terms into the respective fields and click the **Search** button. Button **Clear** will reset all search parameters.

Acquisition ID:

Workflow ID:

Crawler Name:

Status:

Group:

Start Date

From: Select a date

To: Select a date

Last Update Date

From: Select a date

To: Select a date

Acquisitions list 164 **result(s) found**

Acquisition ID	Workflow ID
743575	○ 2e049f3e-b8be-41c5-9608-720645974209
746502	○ 2e049f3e-b8be-41c5-9608-721041409819
745302	○ 2e049f3e-b8be-41c5-9608-721041418288
745314	○ 2e049f3e-b8be-41c5-9608-721041773866
745305	○ 2e049f3e-b8be-41c5-9608-721041697422
745313	○ 2e049f3e-b8be-41c5-9608-721041769389
743563	○ 2e049f3e-b8be-41c5-9608-720178939181
743024	○ 2e049f3e-b8be-41c5-9608-720596172805
745301	○ 2e049f3e-b8be-41c5-9608-720718044455
746501	○ 2e049f3e-b8be-41c5-9608-721041405722
746505	○ 2e049f3e-b8be-41c5-9608-721041414828
746509	○ 2e049f3e-b8be-41c5-9608-721046023014
745304	○ 2e049f3e-b8be-41c5-9608-721041460831
745311	○ 2e049f3e-b8be-41c5-9608-721041758099
745307	○ 2e049f3e-b8be-41c5-9608-721041712937
745320	○ 2e049f3e-b8be-41c5-9608-721047008752
745319	○ 2e049f3e-b8be-41c5-9608-721047003954
464011	○ 2e049f3e-b8be-41c5-9608-693484647044
743014	○ 2e049f3e-b8be-41c5-9608-720178910396
743020	○ 2e049f3e-b8be-41c5-9608-720551025812
746506	○ 2e049f3e-b8be-41c5-9608-721041415986
745318	○ 2e049f3e-b8be-41c5-9608-721046171212

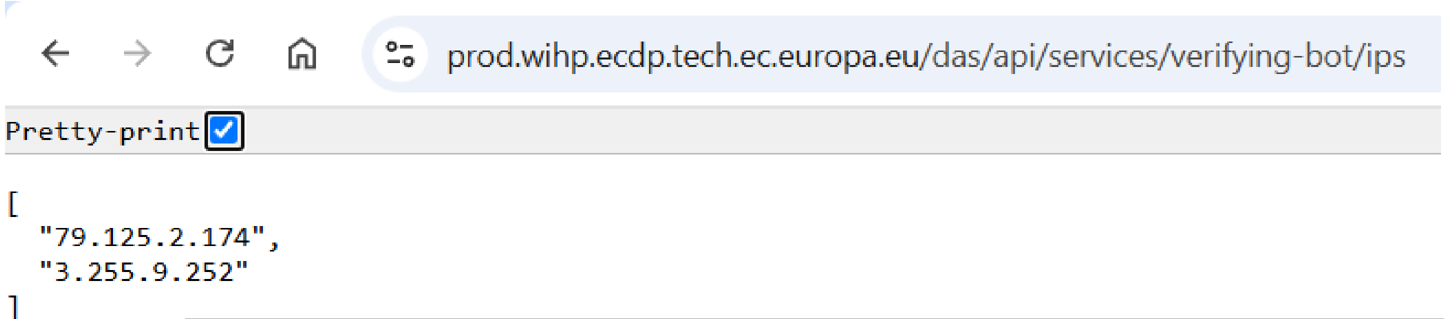
Continuous improvement of the DAS

Version 7 was released in 2023 Oct

- 2nd Security testing and fixing issues
- Handling large number of sources for the OBEC use case
- Moving from Elastic to OpenSearch
- Automatic stopping of acquisitions
- Verifying bot IPs

Version 8 was released in 2024 May

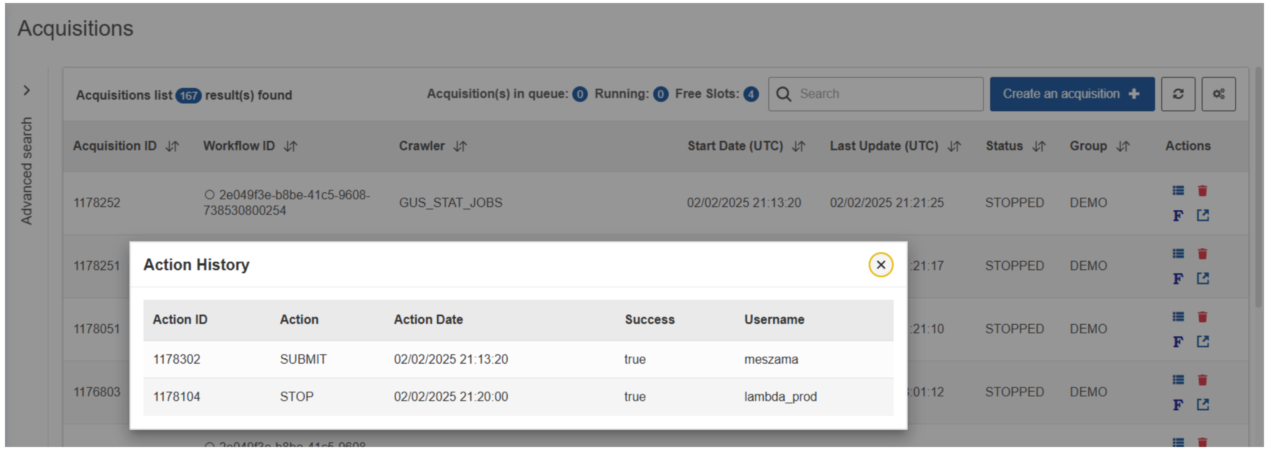
- 3rd Security testing and update
- Exporting data to S3 and access it through API
- Scheduling of acquisition
- Automatic resizing of resources



prod.wihp.ecdp.tech.ec.europa.eu/das/api/services/verifying-bot/ips

Pretty-print ☑

```
[
  "79.125.2.174",
  "3.255.9.252"
]
```



Acquisitions

Acquisitions list 167 result(s) found   Acquisition(s) in queue: 0 Running: 0 Free Slots: 4

Acquisition ID	Workflow ID	Crawler	Start Date (UTC)	Last Update (UTC)	Status	Group	Actions
1178252	○ 2e049f3e-b8be-41c5-9608-738530800254	GUS_STAT_JOBS	02/02/2025 21:13:20	02/02/2025 21:21:25	STOPPED	DEMO	F
1178251				:21:17	STOPPED	DEMO	F
1178051				:21:10	STOPPED	DEMO	F
1176803				:01:12	STOPPED	DEMO	F
	○ 2e049f3e-b8be-41c5-9608						

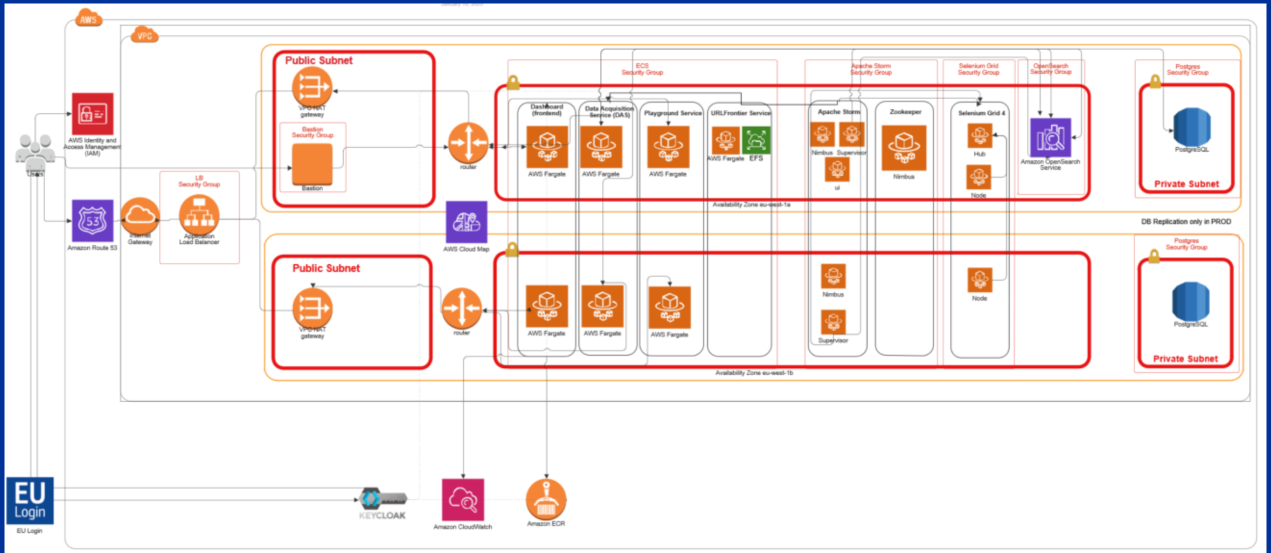
Action History ⊗

Action ID	Action	Action Date	Success	Username
1178302	SUBMIT	02/02/2025 21:13:20	true	meszama
1178104	STOP	02/02/2025 21:20:00	true	lambda_prod

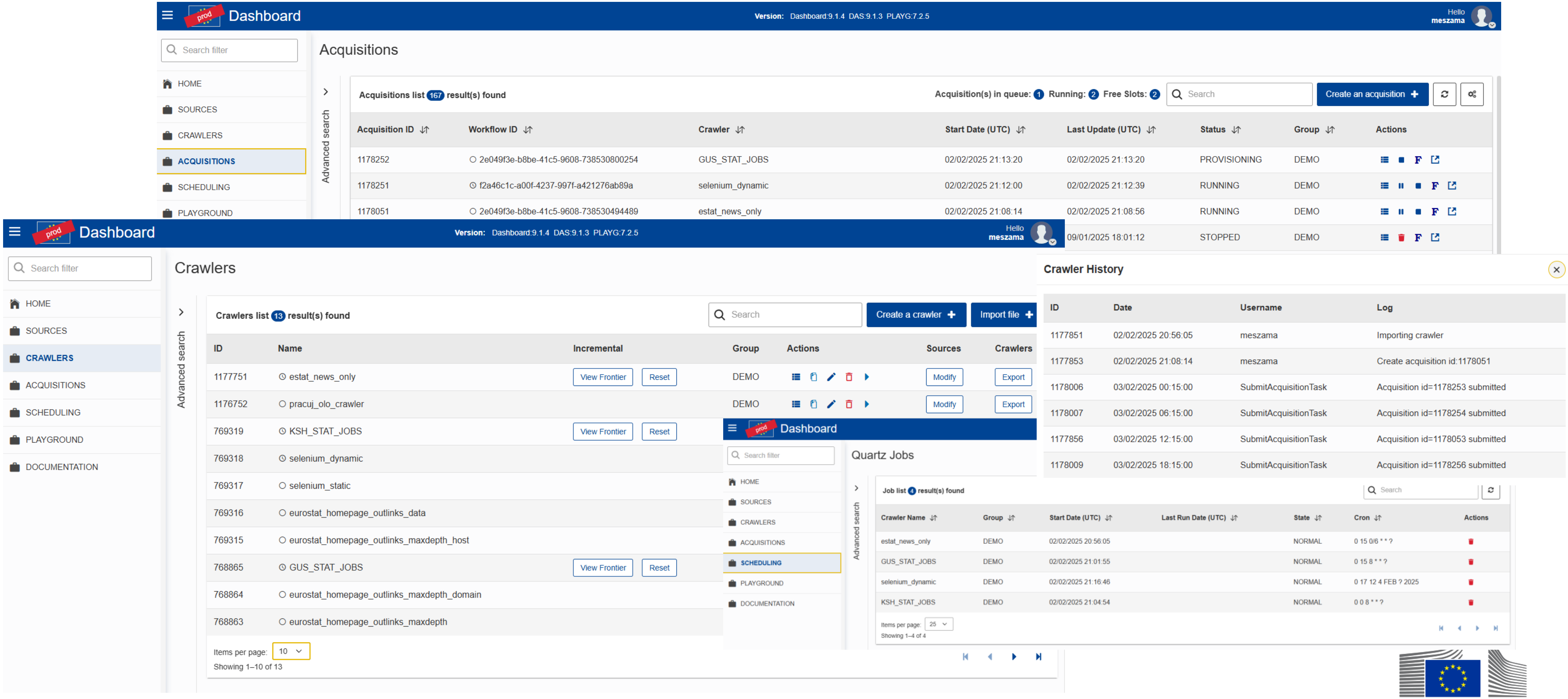
Latest version

Version 9 was released in 2025 Jan

- Introduction of URLfrontier and incremental crawling
- Improved scheduling
- Availability of crawler history
- Possibility to store only extracted text without source html
- No parallel run of the same crawler



Current state



Version: Dashboard:9.1.4 DAS:9.1.3 PLAYG:7.2.5

Acquisitions

Acquisitions list 167 result(s) found

Acquisition(s) in queue: 1 Running: 2 Free Slots: 2

Acquisition ID	Workflow ID	Crawler	Start Date (UTC)	Last Update (UTC)	Status	Group	Actions
1178252	○ 2e049f3e-b8be-41c5-9608-738530800254	GUS_STAT_JOBS	02/02/2025 21:13:20	02/02/2025 21:13:20	PROVISIONING	DEMO	
1178251	○ f2a46c1c-a00f-4237-997f-a421276ab89a	selenium_dynamic	02/02/2025 21:12:00	02/02/2025 21:12:39	RUNNING	DEMO	
1178051	○ 2e049f3e-b8be-41c5-9608-738530494489	estat_news_only	02/02/2025 21:08:14	02/02/2025 21:08:56	RUNNING	DEMO	
				09/01/2025 18:01:12	STOPPED	DEMO	

Crawlers

Crawlers list 13 result(s) found

ID	Name	Incremental	Group	Actions	Sources	Crawlers
1177751	○ estat_news_only	View Frontier Reset	DEMO		Modify	Export
1176752	○ pracuj_olo_crawler		DEMO		Modify	Export
769319	○ KSH_STAT_JOBS	View Frontier Reset				
769318	○ selenium_dynamic					
769317	○ selenium_static					
769316	○ eurostat_homepage_outlinks_data					
769315	○ eurostat_homepage_outlinks_maxdepth_host					
768865	○ GUS_STAT_JOBS	View Frontier Reset				
768864	○ eurostat_homepage_outlinks_maxdepth_domain					
768863	○ eurostat_homepage_outlinks_maxdepth					

Items per page: 10
Showing 1–10 of 13

Crawler History

ID	Date	Username	Log
1177851	02/02/2025 20:56:05	meszama	Importing crawler
1177853	02/02/2025 21:08:14	meszama	Create acquisition id:1178051
1178006	03/02/2025 00:15:00	SubmitAcquisitionTask	Acquisition id=1178253 submitted
1178007	03/02/2025 06:15:00	SubmitAcquisitionTask	Acquisition id=1178254 submitted
1177856	03/02/2025 12:15:00	SubmitAcquisitionTask	Acquisition id=1178053 submitted
1178009	03/02/2025 18:15:00	SubmitAcquisitionTask	Acquisition id=1178256 submitted

Quartz Jobs

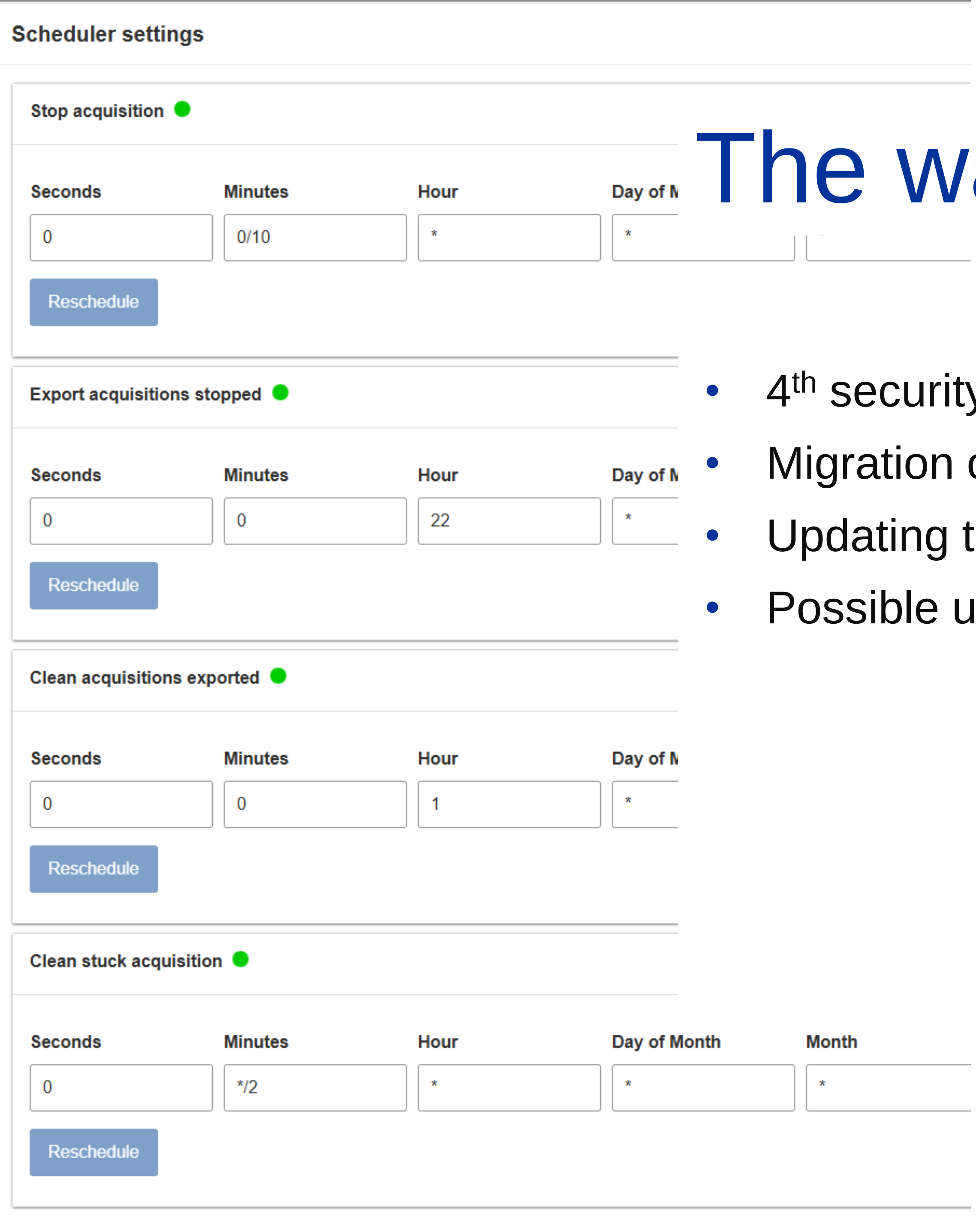
Job list 4 result(s) found

Crawler Name	Group	Start Date (UTC)	Last Run Date (UTC)	State	Cron	Actions
estat_news_only	DEMO	02/02/2025 20:56:05		NORMAL	0 15 0/6 * * ?	
GUS_STAT_JOBS	DEMO	02/02/2025 21:01:55		NORMAL	0 15 8 * * ?	
selenium_dynamic	DEMO	02/02/2025 21:16:46		NORMAL	0 17 12 4 FEB ? 2025	
KSH_STAT_JOBS	DEMO	02/02/2025 21:04:54		NORMAL	0 0 8 * * ?	

Items per page: 25
Showing 1–4 of 4

The way forward

- 4th security testing and corrections
- Migration of OJA-DPS crawlers on the DAS
- Updating the components
- Possible use by the new individual grants



Scheduler settings

Stop acquisition ●

Seconds	Minutes	Hour	Day of M
0	0/10	*	*

Reschedule

Export acquisitions stopped ●

Seconds	Minutes	Hour	Day of M
0	0	22	*

Reschedule

Clean acquisitions exported ●

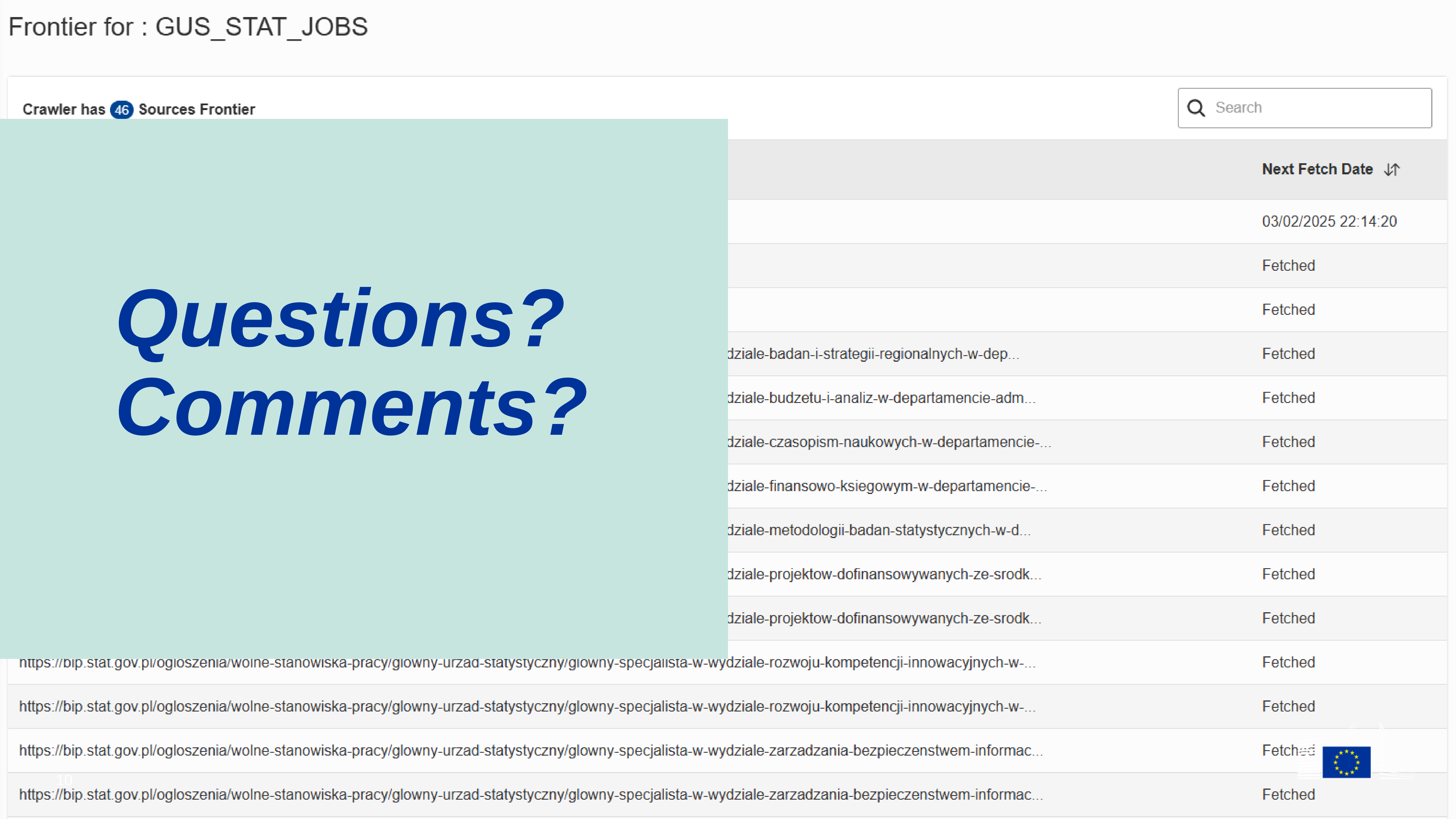
Seconds	Minutes	Hour	Day of M
0	0	1	*

Reschedule

Clean stuck acquisition ●

Seconds	Minutes	Hour	Day of Month	Month
0	*/2	*	*	*

Reschedule



Questions? Comments?

Frontier for : GUS_STAT_JOBS

Crawler has 46 Sources Frontier

	Next Fetch Date
	03/02/2025 22:14:20
	Fetched
	Fetched
dziale-badan-i-strategii-regionalnych-w-dep...	Fetched
dziale-budzetu-i-analiz-w-departamencie-adm...	Fetched
dziale-czasopism-naukowych-w-departamencie-...	Fetched
dziale-finansowo-ksiegowym-w-departamencie-...	Fetched
dziale-metodologii-badan-statystycznych-w-d...	Fetched
dziale-projektow-dofinansowywanych-ze-srodk...	Fetched
dziale-projektow-dofinansowywanych-ze-srodk...	Fetched
https://bip.stat.gov.pl/ogloszenia/wolne-stanowiska-pracy/glowny-urzad-statystyczny/glowny-specjalista-w-wydziale-rozwoju-kompetencji-innowacyjnych-w-...	Fetched
https://bip.stat.gov.pl/ogloszenia/wolne-stanowiska-pracy/glowny-urzad-statystyczny/glowny-specjalista-w-wydziale-rozwoju-kompetencji-innowacyjnych-w-...	Fetched
https://bip.stat.gov.pl/ogloszenia/wolne-stanowiska-pracy/glowny-urzad-statystyczny/glowny-specjalista-w-wydziale-zarzadzania-bezpieczenstwem-informac...	Fetched
https://bip.stat.gov.pl/ogloszenia/wolne-stanowiska-pracy/glowny-urzad-statystyczny/glowny-specjalista-w-wydziale-zarzadzania-bezpieczenstwem-informac...	Fetched

Thank you