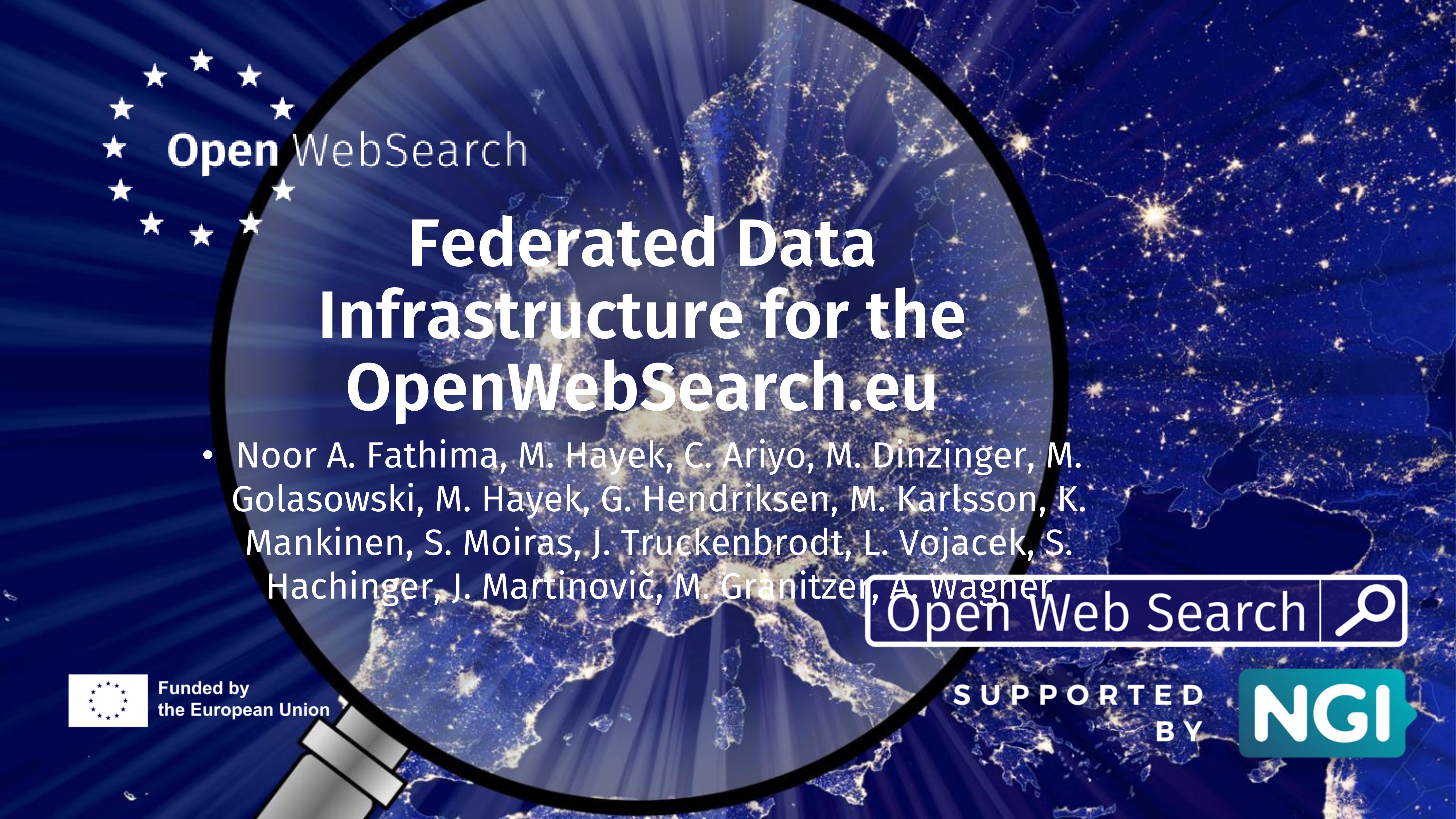


Open WebSearch

Federated Data Infrastructure for the OpenWebSearch.eu

- Noor A. Fathima, M. Hayek, C. Ariyo, M. Dinzinger, M. Golasowski, M. Hayek, G. Hendriksen, M. Karlsson, K. Mankinen, S. Moiras, J. Truckenbrodt, L. Vojacek, S. Hachinger, J. Martinovič, M. Granitzer, A. Wagner

Open Web Search

Funded by
the European Union

SUPPORTED BY NGI

OpenWebSearch.eu Project (OWS)

➢ **Horizon Europe Project**

- Duration: 3 years (plus 6 months?) 2022 – 2025/26
- Collaborative effort with 14 research institutions, leveraging both cloud and high-performance computing (HPC).

➢**Vision**

- Restore an open search engine market as a basis for a new Internet Search
- Aims to create a scalable and open European web search infrastructure.
- Empower Europe’s researchers, innovators and businesses to systematically tap into the Web as business and innovation resource.
- Main output of the project: Open Web Index

Objective - Federated Data Infrastructure

- → Infrastructure partners addresses technology stacks and development for a joint distributed storage and compute infrastructure.
- → Confluence of High-Performance Computing (HPC) resources, and Infrastructure-as-a-Service cloud (IaaS-cloud) solutions.
- → Duration: M1 - M36
- → Goals:
  - Provide state-of-the-art storage and compute infrastructure to be used by the technical WP1-4 as backbone for development and sustainable hosting of services and data.
  - Develop and manage highly scalable, reliable and secure computing infrastructures
  - Run core services and store core data products

14 Partners plus Third Party partners

Open WebSearch .eu

UNIVERSITÄT PASSAU

Webis.de

UNIVERSITÄT LEIPZIG

Bauhaus-Universität Weimar

Radboud Universiteit

DLR

TU Graz

Research

lrz

CERN

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

CSC ICT Solutions for Brilliant Minds

Infrastructure

osf open search foundation

nlnet FOUNDATION

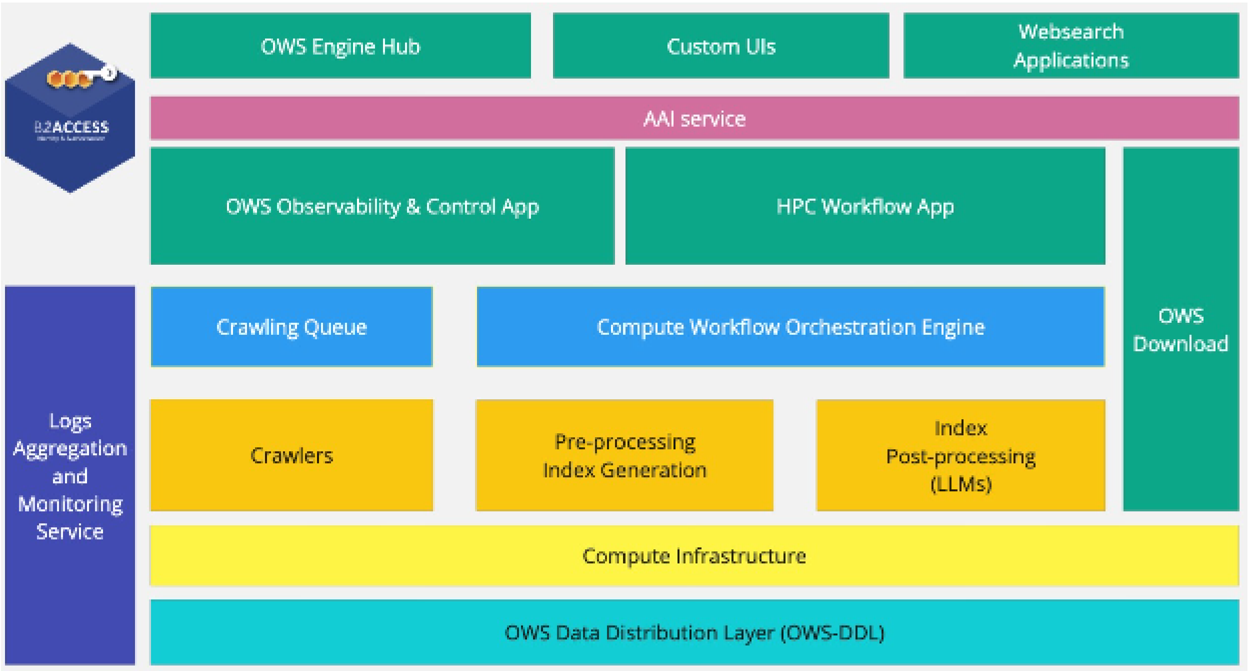
suma-ev

NGOs

A1

Businesses

Pilot OWS Federated Data Infrastructure (OWS-FDI)



B2ACCESS

OWS Engine Hub

Custom UIs

Websearch Applications

AAI service

OWS Observability & Control App

HPC Workflow App

OWS Download

Logs Aggregation and Monitoring Service

Crawling Queue

Compute Workflow Orchestration Engine

Crawlers

Pre-processing Index Generation

Index Post-processing (LLMs)

Compute Infrastructure

OWS Data Distribution Layer (OWS-DDL)

Layer 1 (green boxes) are collectively referred to as ***Interfaces.***

Layer 2, (purple box), is dedicated to Authentication and Authorization Infrastructure layer, also known as the ***AAI layer***.

Layer 3 (orange boxes) includes the ***Crawling, Pre-processing, and Index Generation layer***

Layer 4, (yellow box), constitutes the ***Compute Infrastructure*** **layer** necessary for data processing tasks.

The Data Distribution Layer, visualized in a cyan box (bottom most), handles the dissemination of data across the network.

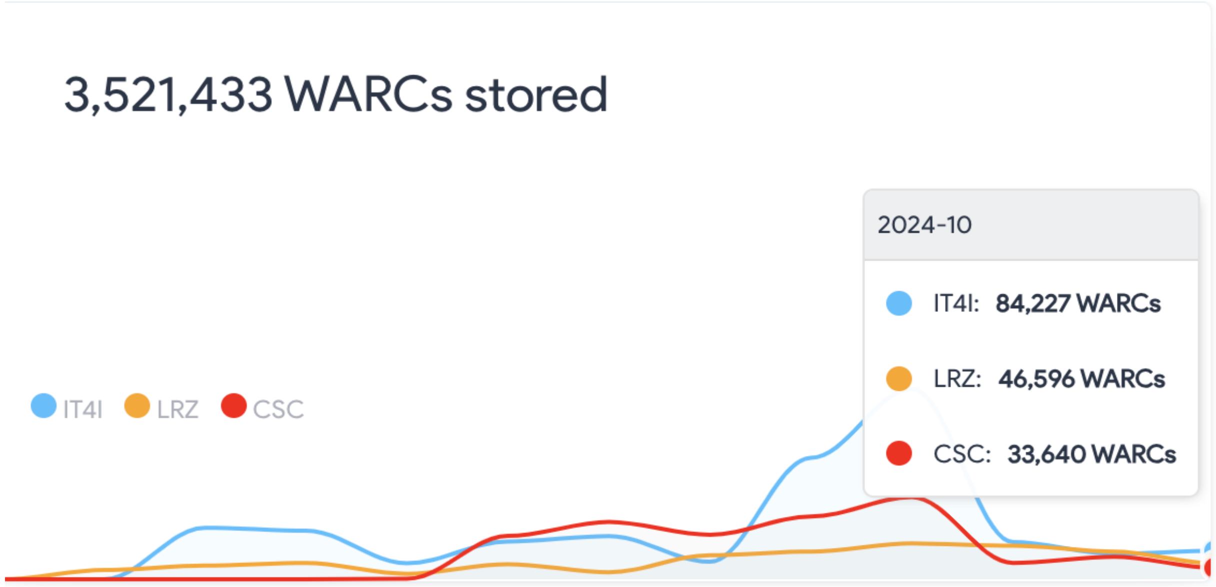
Supplementing these layers are the

Single Sign-On (SSO) feature, which integrates with nearly all components to streamline user authentication and system access.

Logs Aggregation and Monitoring Service which overlays on most components.

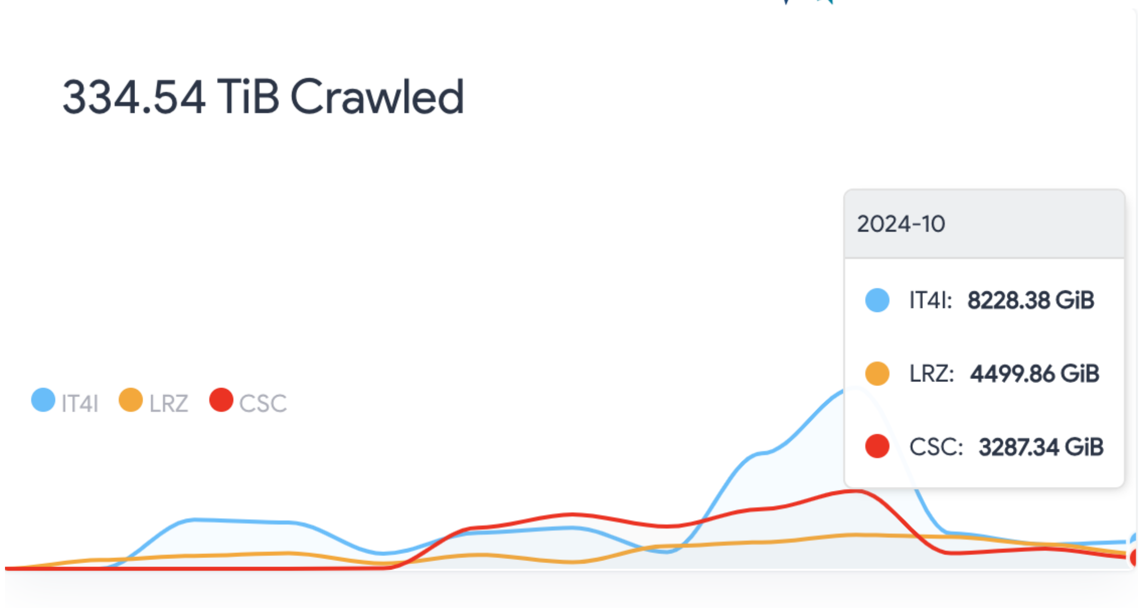
The two backend components ***Crawling Queue and Compute workflow orchestration** engine.*

Current Status


3,521,433 WARCs stored

2024-10

- IT4I: 84,227 WARCs
- LRZ: 46,596 WARCs
- CSC: 33,640 WARCs


334.54 TiB Crawled

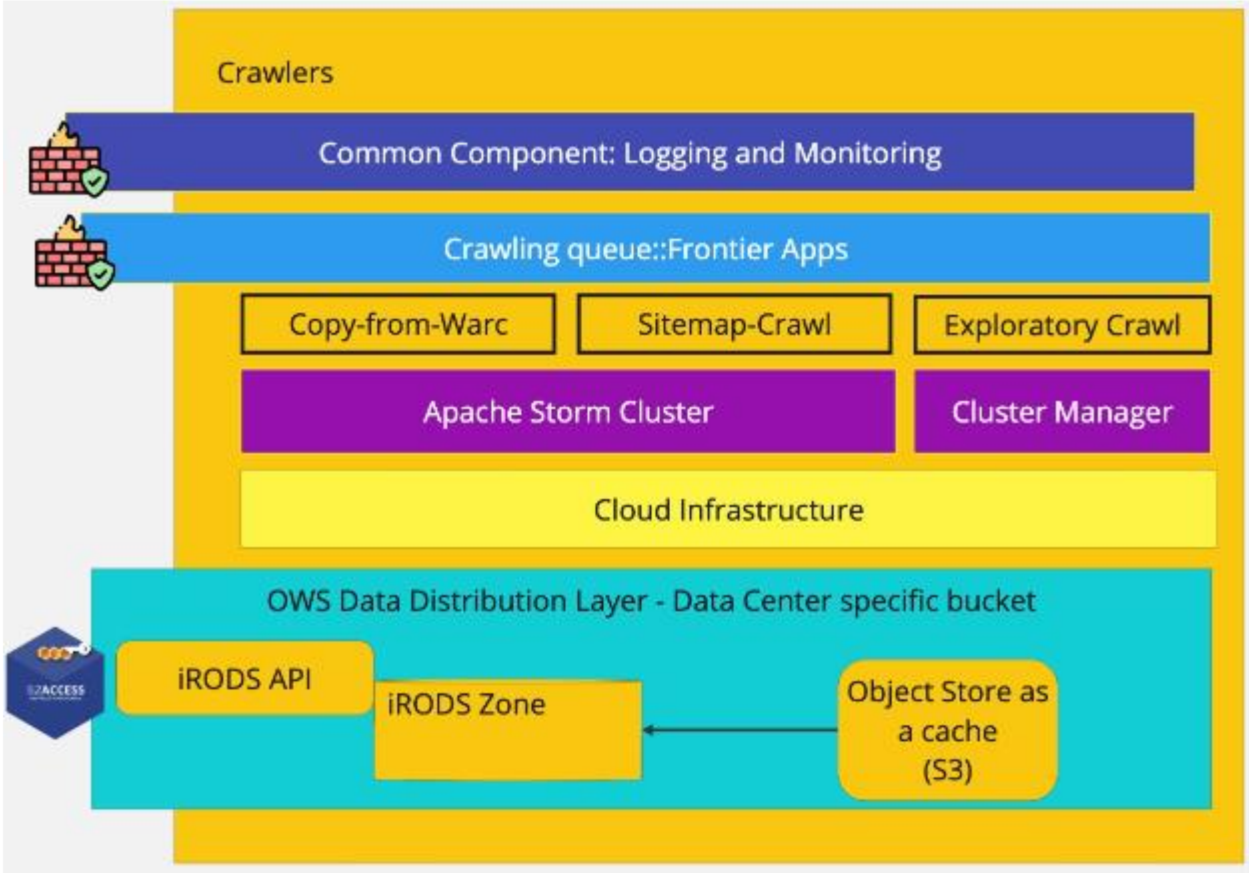
2024-10

- IT4I: 8228.38 GiB
- LRZ: 4499.86 GiB
- CSC: 3287.34 GiB

Statistics on General-Purpose crawling (since August 2023) - Status August 2024

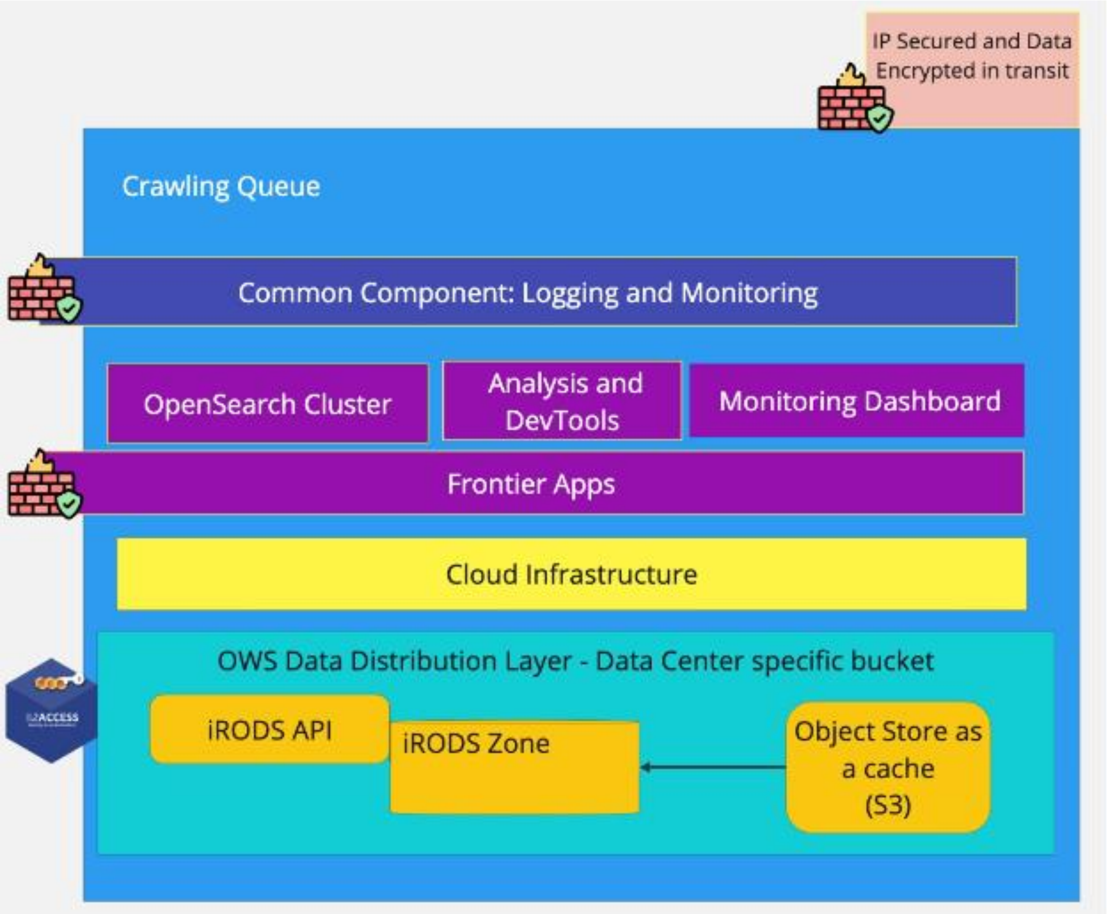
Visited URLs (per day) Note that this number has been achieved during a test crawl spanning 20 days (20.06.2024 - 11.07.2024).	Up to ~105M / day
Visited URLs (total)	10,757,755,422 until 6th of August 2024
Unique visited URLs	1,170,925,504 until 6th of May 2024
Unique visited hosts	40,531,574 until 6th of May 2024
Number of unique languages	184

Crawlers


Crawlers
Common Component: Logging and Monitoring
Crawling queue::Frontier Apps
Copy-from-Warc
Sitemap-Crawl
Exploratory Crawl
Apache Storm Cluster
Cluster Manager
Cloud Infrastructure
OWS Data Distribution Layer - Data Center specific bucket
iRODS API
iRODS Zone
Object Store as a cache (S3)

- → OWler crawls the internet to gather data, generating WARC files.
- → Storage of WARC files in the OWS-DDL object store for shared access.
- → Crawlers operational at IT4I and LRZ‘s cloud infrastructure.
- → Interacts with the crawler queue via Frontier Apps interface both at CERN.
- → Connects with Logging and Monitoring server to provide logs & metrics data.

Crawling Queue:



IP Secured and Data Encrypted in transit
Crawling Queue
Common Component: Logging and Monitoring
OpenSearch Cluster
Analysis and DevTools
Monitoring Dashboard
Frontier Apps
Cloud Infrastructure
OWS Data Distribution Layer - Data Center specific bucket
iRODS API
iRODS Zone
Object Store as a cache (S3)
ACCESS

- → Manages and monitors the status of crawled and to-be-crawled URLs.
- → Frontier – data structure for storing URLs discovered/visited during crawl
- → OpenSearch backend for Frontier apps which interface via URLFrontier API with the crawler nodes.

Integration with OWS-DDL:

- → Accesses different parts of the OWI and intermediary files like transfer logs and public metrics.

Crawling Queue:

Operational Setup:

→ Deployed at CERN on cloud infrastructure, accessible via SSH.

→ OpenSearch leads authentication, maintaining an internal database of user roles and hashed passwords.

→ CERN manage basic-auth credentials.

Expansion and Scalability:

→ Expanding data nodes and implementing warm/hot storage for scalability.

→ Optimizing memory usage by adhering to maximum standard limits per process.

Future Development Goals:

→ Connecting to OWS-DDL for displaying public logs and metrics in the main OWS-O&C App.

Crawling Queue:

Operational Setup:

→ Deployed at CERN on cloud infrastructure, accessible via SSH.

→ OpenSearch leads authentication, maintaining an internal database of user roles and hashed passwords.

→ CERN manage basic-auth credentials.

Expansion and Scalability:

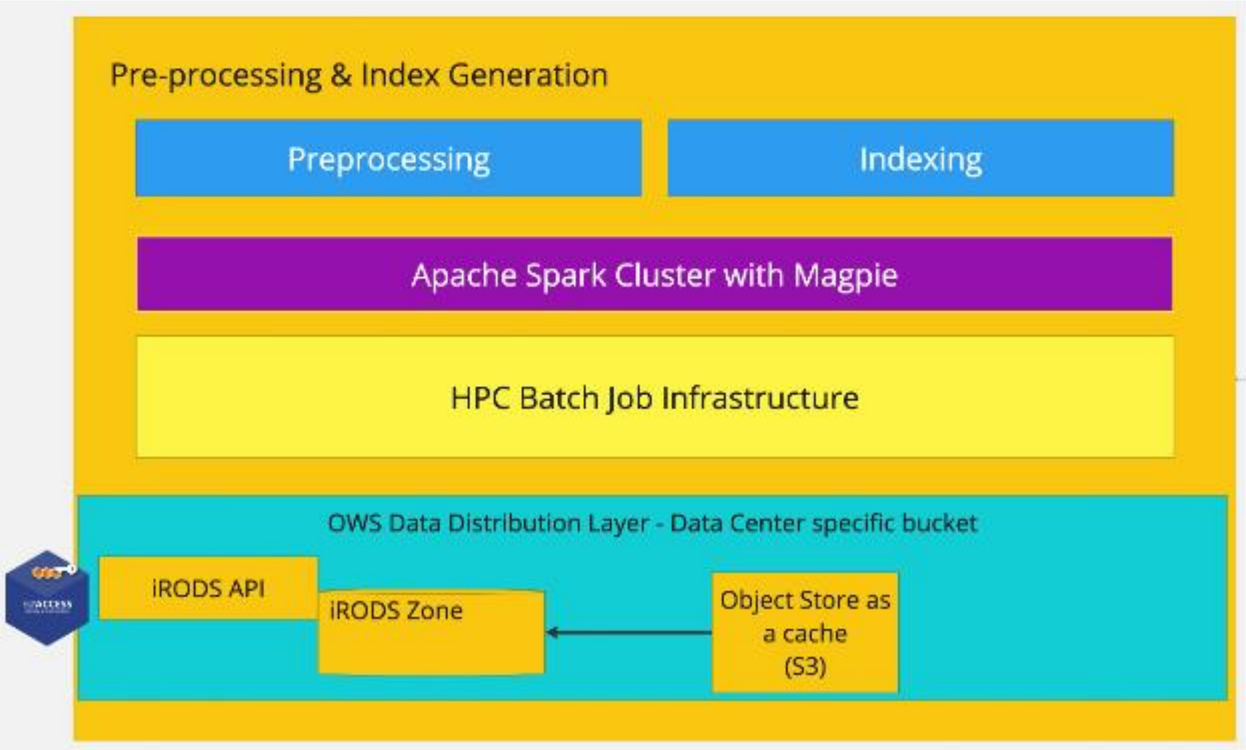
→ Expanding data nodes and implementing warm/hot storage for scalability.

→ Optimizing memory usage by adhering to maximum standard limits per process.

Future Development Goals:

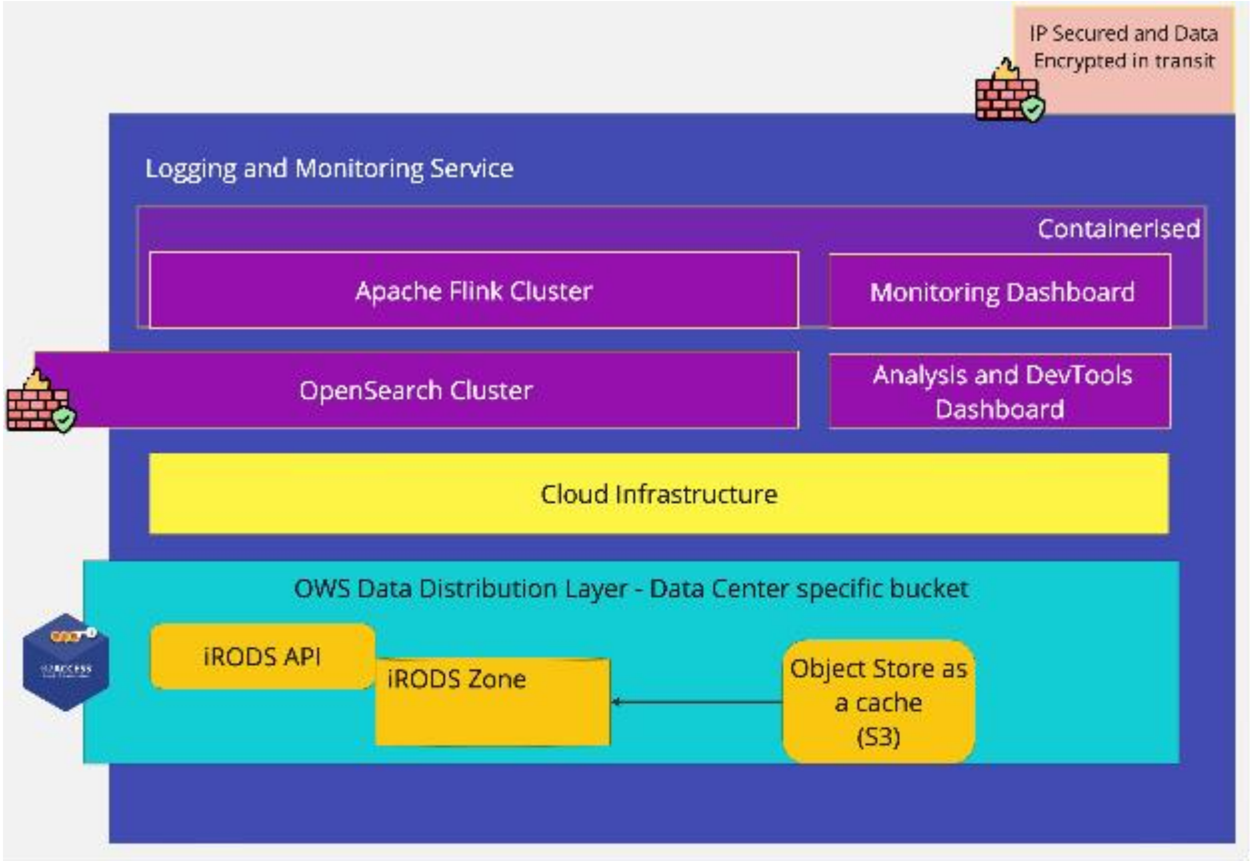
→ Connecting to OWS-DDL for displaying public logs and metrics in the main OWS-O&C App.

Pre-processing and Index Generation


Pre-processing & Index Generation
Preprocessing
Indexing
Apache Spark Cluster with Magpie
HPC Batch Job Infrastructure
OWS Data Distribution Layer - Data Center specific bucket
iRODS API
iRODS Zone
Object Store as a cache (S3)

- → Crawled data subjected to a pre-processing stage before transforming it into an indexed format.
- → Utilizing Apache Spark for batch processing.
- → Running at IT4I and LRZ - using HPC infrastructure.
- → Originally employed Magpie script collection for deploying Spark cluster.
- → Now integrated with HPC Work-flow App and Compute Work-flow Orchestration Engine.
- → TIRA platform instance hosted at CSC.

Logging and Monitoring service

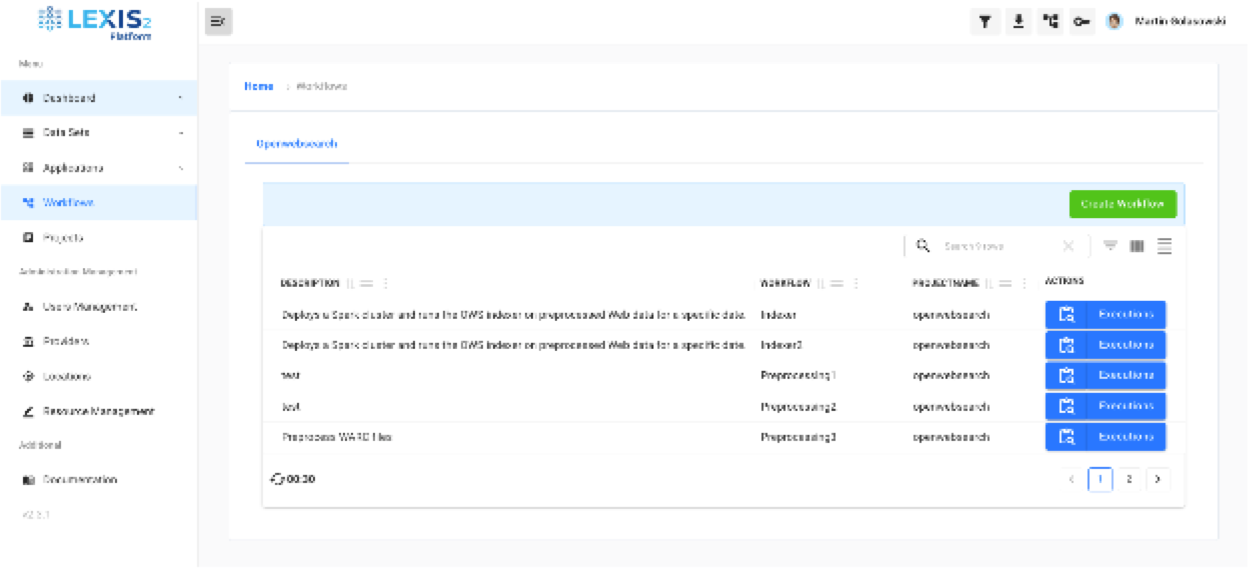


IP Secured and Data Encrypted in transit
Logging and Monitoring Service
Containerised
Apache Flink Cluster
Monitoring Dashboard
OpenSearch Cluster
Analysis and DevTools Dashboard
Cloud Infrastructure
OWS Data Distribution Layer - Data Center specific bucket
iRODS API
iRODS Zone
Object Store as a cache (S3)

- → Designed to gather logs and metrics from all components.
- → Interfaces with the OWS Observability & Control App, ensuring accessibility of public log data and metrics.
- → Includes cron-jobs for maintaining the Blacklist index.

HPC Workflow App & Compute workflow and orchestration engine

LEXIS2 Platform



DESCRIPTION	WORKFLOW	PROJECTNAME	ACTIONS
Deploys a Spark cluster and runs the OWS indexer on preprocessed Web data for a specific date.	Indexer	openwebsearch	Executions
Deploys a Spark cluster and runs the OWS indexer on preprocessed Web data for a specific date.	Indexer2	openwebsearch	Executions
test	Preprocessing1	openwebsearch	Executions
test	Preprocessing2	openwebsearch	Executions
Preprocess WARC files	Preprocessing3	openwebsearch	Executions

→ This GUI facilitates efficient management and streamlining of WP2 and WP3 workflows.

→ Enables creation and management of workflow executions, utilizing scripts and DAGs (Directed Acyclic Graphs).

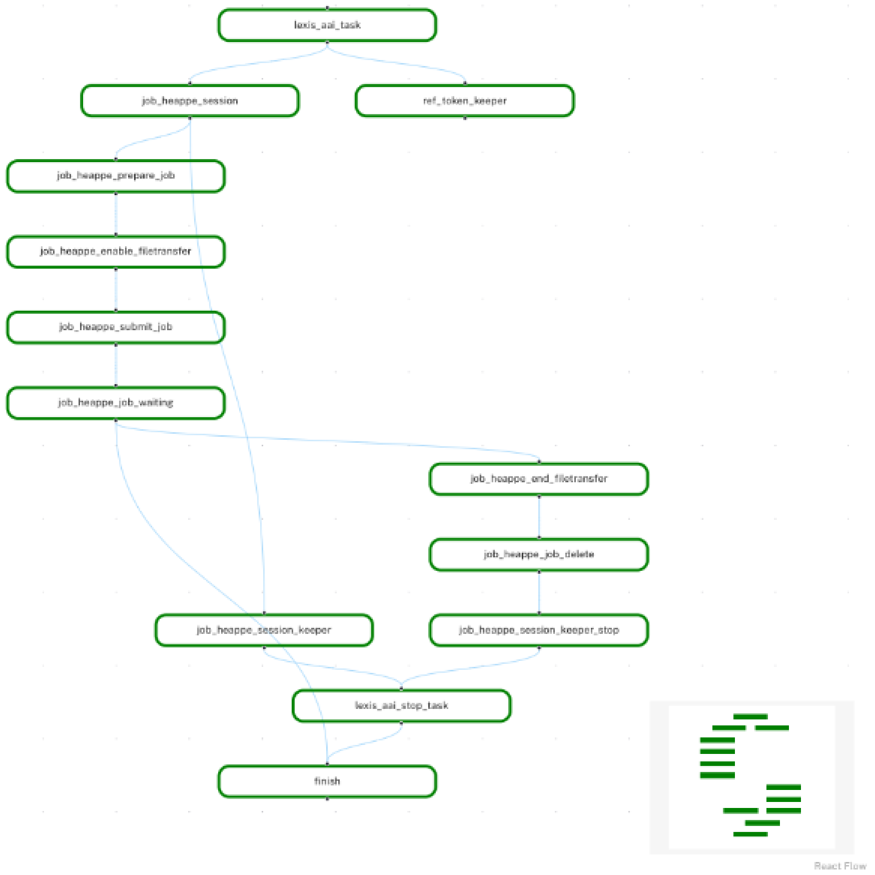
Dataset Management Capabilities

→ Allows creation, viewing, and modification of datasets, including metadata handling.

→ Facilitates file upload, download, and deletion within dataset directory structures.

→ Integrates datasets with OWS-DDL for data discoverability and replication and makes them available for download.

HPC Workflow App & Compute workflow and orchestration engine



→ Users can specify execution resources like CPUs.

→ LEXIS framework offers supercomputing and cloud resources access.

Apache Spark Cluster Deployment

→ Configuring and deploying Apache Spark clusters within the HPC environment.

→ Submitting Spark jobs for indexing preprocessed data.

→ Progressing towards integration with the data staging and datasets API provided by LEXIS.

Authentication and Authorization Infrastructure

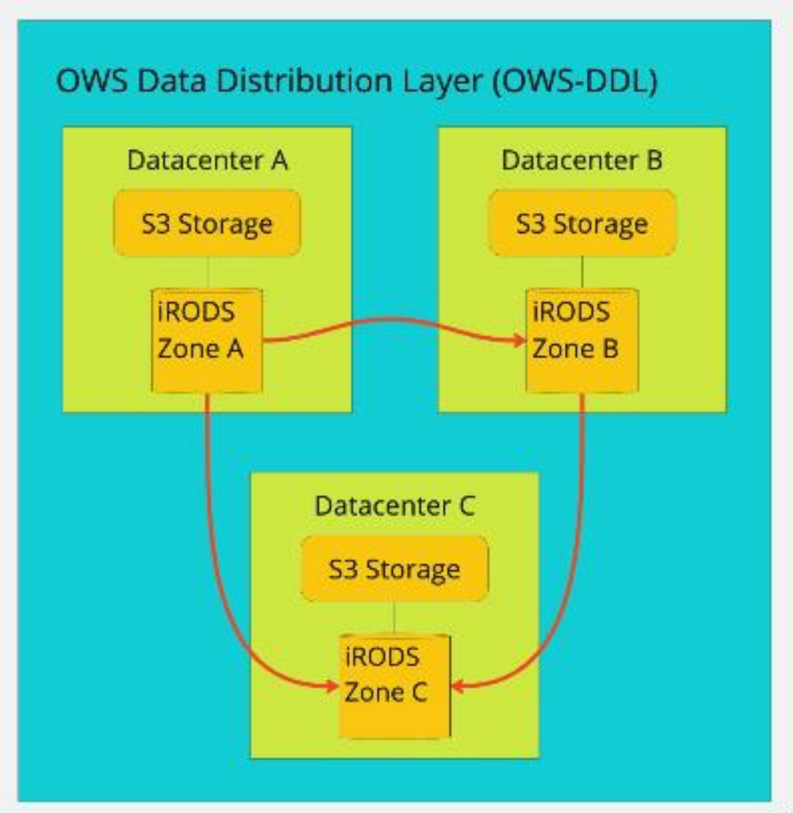
B2ACCESS
Identity & Authorisation

- → Seamless access across diverse services and administrative domains in OWS-FDI.
- → Utilizes Single Sign-On (SSO) through dedicated Keycloak for simplified user experience.
- → LEXIS is also integrated
- → iRODS zones integrated with LEXIS Keycloak, direct access possible

→ **Questions:**

- Do we need personal/citizen ID level of assurance? (eIDAS, national/bank identities)
- Or do we need to open more? Allow all, keep only logs?
- Standalone iRODS, Keycloak and B2A integration?

OWS Data Distribution Layer (OWS-DDL)



→ Utilizes geo-distributed storage and mirroring for data redundancy and safety.

→ Incorporates iRODS and EUDAT-B2SAFE for data management.

Multi-Site Data Storage and Management

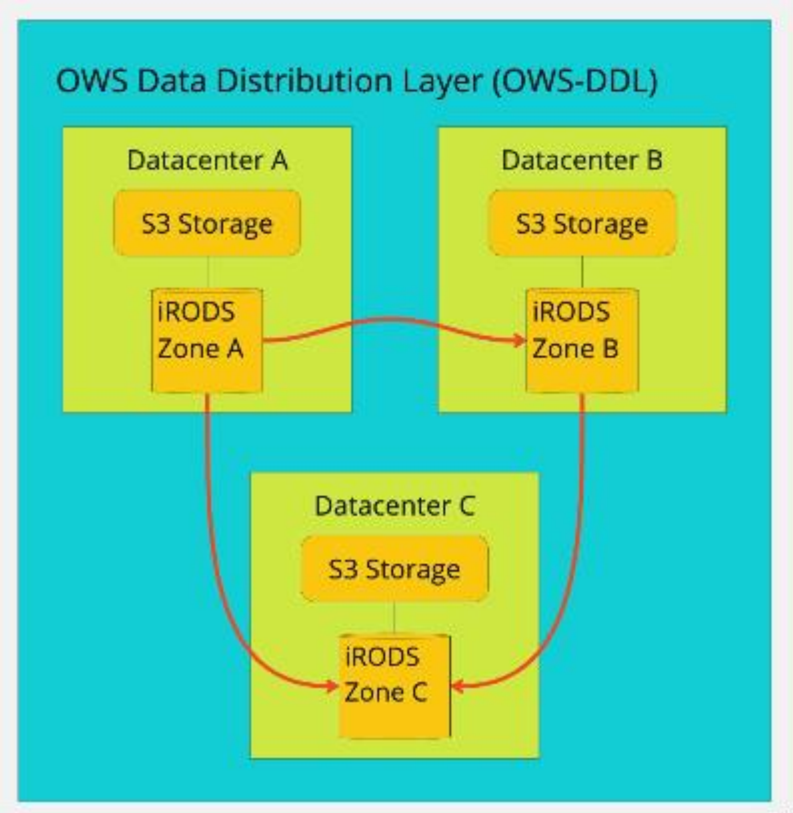
→ Data management distributed across all 5 data centers.

→ S3 Used as temporary storage, iRODS for publishing and final datasets

→ All iRODS zones interconnected for data transfer and sharing.

→ Data is published through the **LEXIS Platform** using a local iRODS zone.

OWS Data Distribution Layer (OWS-DDL)



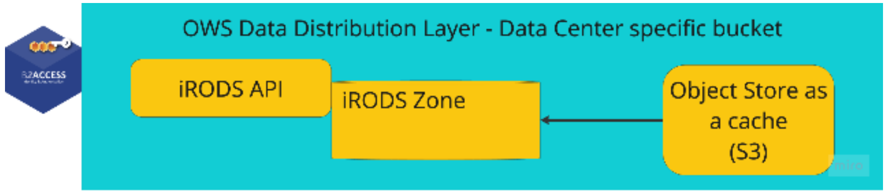
→ Established **distributed data management solution**.

→ Used within **OWS** as an **abstraction of local (POSIX) storage arrays**.

→ Enables **federated data transfer** between connected locations.

→ Data is published through the **LEXIS Platform** using a local iRODS zone.

Integration components of OWS-DDL


OWS Data Distribution Layer - Data Center specific bucket
iRODS API
iRODS Zone
Object Store as a cache (S3)

S3 Protocol:

- Used by **crawlers** and **processing pipelines** for storing **intermediate products** (e.g., raw WARC files).
- **Two options** for S3-compatible storage:
    - Use a **local S3-compatible service**.
    - Deploy a **local MinIO instance** for fast access to processing pipelines.

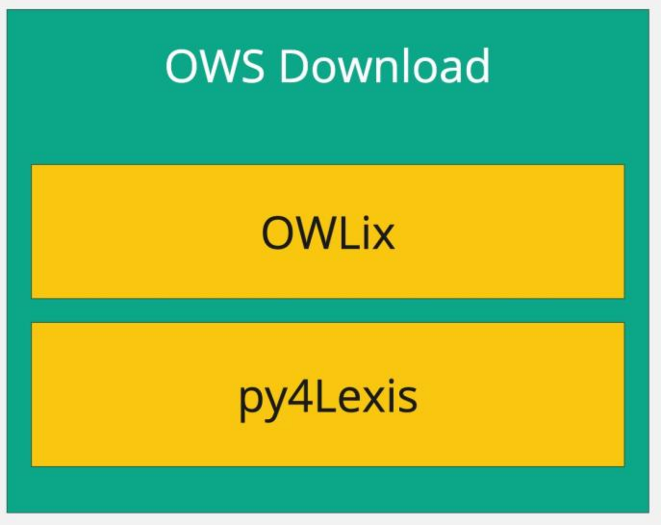
Local S3 Setup:

- **Single service account** with **access and secret key** is preferred for simplicity.
- Variations in **access level control** across S3 implementations can impact setup.

Processing Pipeline Integration:

- Based on **Apache Spark** with **native S3 integration**.
- Supports either **local S3 storage arrays** or **local MinIO instances**.
- Data is processed and then transferred to iRODS for **long-term management**.
- S3 storage serves as a **temporary buffer**, with data removed after a certain time.

Dataset download options



OWS Download

OWLix

py4Lexis

owilix:

Purpose:

- Enhances access to the **Open Web Index (OWI)**.
- Designed for **researchers, developers, and data scientists**.
- Facilitates the **management** and **querying** of web-scale datasets.

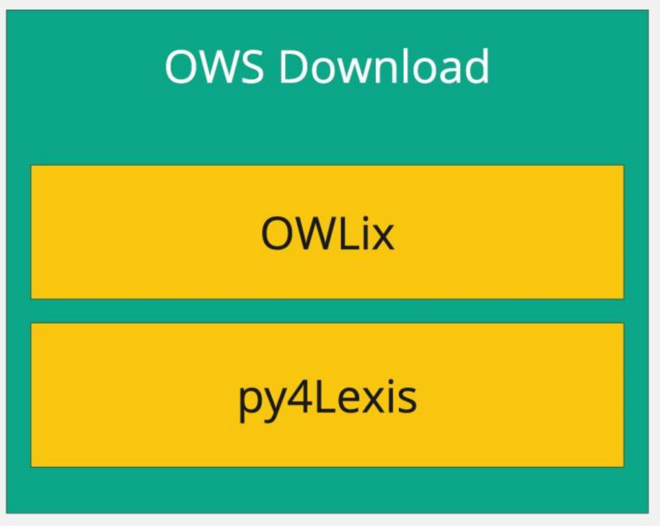
Key Functionalities:

- **Pull**: Retrieval of **OWI-shards.**
- **Push**: Supports **community contributions**.
- **Advanced SQL querying** using **DuckDB** with **Parquet** format.
- Efficiently manages both **local** and **remote datasets**.

Architecture:

- Integrates with the broader **OWI ecosystem** using:
- **iRODS** for parallel downloads.
- **py4lexis** for general data access.
- Supports **secure authentication** with token-based access (valid for five days).

Dataset download options


OWS Download

OWLix

py4Lexis

py4lexis:

Purpose:

- Acts as a **client library** for interacting with the **LEXIS Platform**.
- Designed for managing **large datasets** in a **distributed** environment.

Key Functionalities:

- Supports **data upload**, **staging**, **compression**, **encryption**, and **metadata management**.
- **Automates handling** of distributed datasets across multiple storage nodes.
- Abstracts complex **orchestration tasks** to simplify user operations.

Integration:

- Interfaces with storage systems like **iRODS** for:
  - **Efficient data transfer**.
  - **Secure access**.
  - **Workflow automation**.
- Supports high-performance computing and data management environments.

Authentication Process:

- Users log in via the **LEXIS login page**.
- **B2Access credentials** can be used for secure access to the **OWS-DDL**.

Future:

- → Migration of security model of Frontier tier.
- → Benchmarking the Frontier Tier to optimise the scaled process.
- → Scaling up pre-processing/enrichment and indexing activities.
- → Stabilise the download operations

Questions?

Open WebSearch .eu