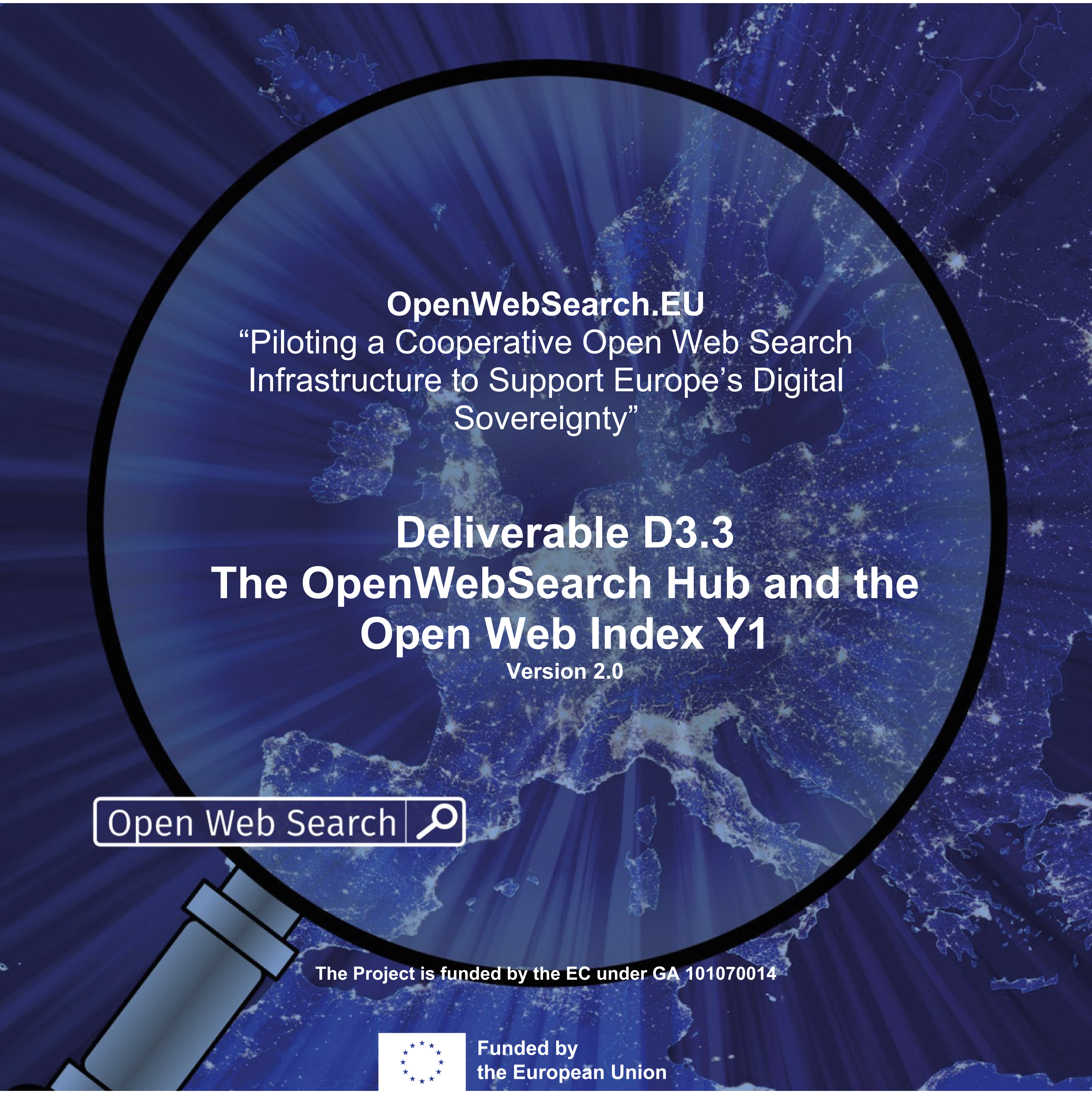


OpenWebSearch.EU
“Piloting a Cooperative Open Web Search Infrastructure to Support Europe’s Digital Sovereignty”

Deliverable D3.3
The OpenWebSearch Hub and the Open Web Index Y1

Version 2.0

The Project is funded by the EC under GA 101070014

Funded by
the European Union

OPENWEBSEARCH.EU

Preliminaries

i. Project Info

Project number	101070014
Project acronym	OWS.eu
Project name	OpenWebSearch.eu – Piloting a Cooperative Open Web Search Infrastructure to Support Europe's Digital Sovereignty
Call	HORIZON-CL4-2021-HUMAN-01
Topic	HORIZON-CL4-2021-HUMAN-01-05
Type of action	HORIZON-RIA
Responsible unit	DG CNECT
Project starting date / Duration	01/09/2022 – 08/2025 (36 months)
Project reporting period	1
Project Coordinator	Prof. Dr. Michael Granitzer, University of Passau

ii. Project Partners

Acronym	Partner
UNI PASSAU	University of Passau
BADW-LRZ	Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities
RU	Radboud University
WEBIS	Leipzig University
TUGraz	Graz University of Technology
DLR	German Aerospace Center
CERN	CERN
IT4I@VSB	VSB - TECHNICAL UNIVERSITY OF OSTRAVA
BUW	Bauhaus-Universität Weimar (Associated Partner)
OSF	Open Search Foundation e.V.
A1	A1 Slovenia

SUMA-EV	Suma e.V. (Associated Partner)
CSC	CSC – Tieteen tietotekniikan keskus Oy
NLnet	Stichting NLnet

iii. Deliverable Info

Due Date / Delivery Date	M12
Deliverable Lead	RU
Deliverable type	DEM: Demonstrator, pilot, prototype, plan designs
Dissemination level	PU
Document Status / Version	V2.0
Work-package / Lead Partner	WP3/RU
Main authors	Gijs Hendriksen, Djoerd Hiemstra, Arjen de Vries (WP 3) Sebastian Schmidt, Ines Zelch, Martin Potthast (WP 2) Michael Granitzer, Michael Dinzinger, Saber Zerhoudi (WP 1) Noor Afshan Fathima (WP 5)
Approval	The deliverable expresses the opinion of the authors and has not yet been approved by the EC.

iv. Deliverable Summary

This document describes the first version (out of three) of deliverable D3.3 “The OpenWebSearch Hub and the Open Web Index” of the OpenWebSearch.eu project funded by the EC under the GA 101070014 within the Horizon Europe Framework programme. It specifies the current state of the development and deployment of the Open Web Index (OWI) and the OpenWebSearch Engine Hub (OWSE-HUB), as compared to the proposal phase, and discusses the planning for the next year.

v. Document Management

History of Changes

Name	Version	Publication date	Changes
Initial Version	1.0	22.06.2023	Draft document structure/outline
First Draft	1.1	05.07.2023	First general text
Second Draft	1.2	18.07.2023	Processed feedback and added vision
Semi-final Version	1.3	23.08.2023	Processed feedback from main reviewer
Final Version	2.0	30.08.2023	Processed final feedback

vi. Document Approver(s) and Reviewer(s)

NOTE: All Approvers are required. Records of each approver must be maintained. All reviewers in the list are considered required unless explicitly listed as “Optional”.

Name	Role	Action	Date
Michael Granitzer	Project Coordinator	Review	10/07/23, 23/07/23
Michael Granitzer	WP 1 Lead	Contribution	23/07/23
Christian Gütl	Reviewer	Review	18/08/23

vii. Executive Summary

Two important outputs of the project — and the main focus of work package 3 — are the Open Web Index (OWI) and the OpenWebSearch Engine Hub (OWSE-HUB). Our general vision for both these systems are described in the project proposal (Tasks 3.1 and 3.3). In this deliverable, we make our vision more concrete, and discuss how a federated data architecture (across Europe) can be set up to help achieve this vision.

For the first year, we have put our main focus on the development of a first end-to-end pipeline for building the OWI. In short, this means that WPs 1, 2 and 3 have focused on setting up a basic version of the crawler, preprocessing pipeline and indexer, respectively, with all of these projects running on the infrastructure provided by WP 5. This first version of the full pipeline can be iterated upon in the next months, where we can identify and resolve any issues, and expand or improve functionalities of each part. It also provides us with a good foundation to begin working on the tasks that are due to start in the second year of the project.

Instead of only discussing the outputs of our work (the OWI and OWSE-HUB), we discuss the full technical pipeline we developed this year in detail. This helps to make our progress as transparent as possible and at the same time offers a deeper understanding of the steps we take to crawl the Web and turn these crawls into a usable index.

In summary, we have achieved the following main contributions:

1. End-to-end pipeline: We have developed a full pipeline that crawls, preprocesses and indexes the Web on a federated data infrastructure, which is running successfully at at least one of the infrastructure partners. The output of this pipeline is the Open Web Index.
2. Architecture & Infrastructure: The entire setup operates in cluster tiers, notably: the Crawling Cluster Tier (CCT) for crawling the Web; the Crawler Frontier Tier (CFT) for managing and coordinating the multi-tier crawling process; the Preprocessing and Enrichment Tier (PET) for preprocessing and cleaning of the Web data; and the Indexing and Storage Tier (IST) for indexing and partitioning the data into readily usable index shards.
3. Statistics: As of the end of July, a total of approximately 300M documents has been indexed, spanning over 10.6M unique domains and 179 different languages.
4. Standards: All data outputs are stored and exchanged in well-known, easily usable, standard formats: WARC for crawl outputs; Parquet for metadata stores; and CIFF for indexes.
5. Open-source software: All software developed for these purposes is published as an open-source software package (see Section 4.1). An archived version is also available on Zenodo.

1. Introduction

This document describes our progress on the Open Web Index (OWI) and the OpenWebSearch Engine Hub (OWSE-HUB) after the first year of the project. Before we show the technical details of our work in the first year, we first explain our vision for the OWI and OWSE-HUB in Sections 1.1 and 1.2. This vision includes a federated data structure spread across Europe (see also Figure 1), where each data center is responsible for crawling a set of web pages and cleaning and indexing them locally. After introducing our vision, we present the full technology stack developed thus far to build and deploy the OWI and OWSE-HUB — from crawling and parsing the Web to indexing and deployment.

Rather than focusing on the end product of our efforts (the OWI and eventually the OWSE-HUB), we choose to highlight all the steps required to successfully gather all data and process it into a usable index. This should result in a clear picture of how we aim to build and deploy the OWI and OWSE-HUB, and allow for more transparency into our current progress.

1.1. Vision for the Open Web Index

The Open Web Index (OWI) accommodates a variety of search engines, by making it easy for search engine designers to use (parts of) the index for their own purposes. To this end, we envision the OWI to become a (distributed) information system similar to the well-known Docker hub. Instead of virtual machines, though, the OWI would contain pre-built indexes that would be readily usable. Our general vision for this distributed system is shown in Figure 1.

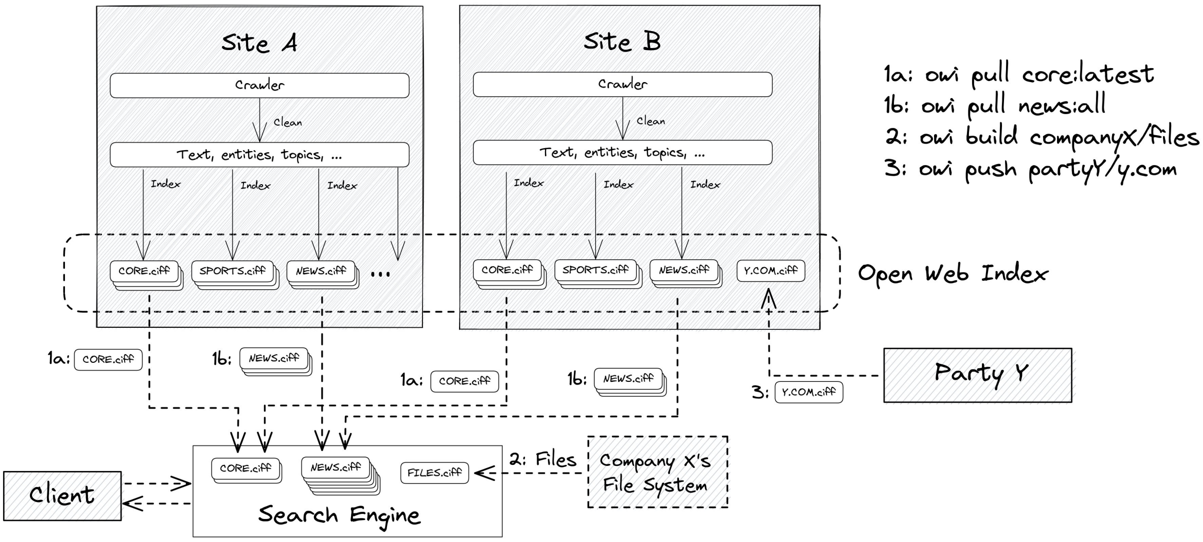


Figure 1: General architecture of the OWI, and the way in which a search engine could interact with the OWI to retrieve (parts of) the index.

Our federated data infrastructure (spread across multiple European data centers) crawls, enriches and indexes web content in a distributed manner. These indexes are fragmented into a set of pre-defined (possibly overlapping) verticals, and continuously updated over time. Aside from these vertical indexes, we also create a “core” index that should contain the most popular and important websites on the Web. Since a small subset of the Web accounts for a large amount of user clicks, a core index containing only this popular subset of the Web should suffice for a large number of basic queries.

When taken together, these index “shards” form the OWI enable a number of downstream uses (also shown in the overview above).

1. Users or organisations can download (or “pull”) a specific, pre-built index.
    a. They can choose a specific timestamp or checkpoint of the index (e.g. “latest” for the most recent version).
    b. They can choose to download a selection of checkpoints, instead of only a single one (e.g. “all” for the complete history of a specific index).
2. Users or organisations can create (or “build”) their own index locally, using a dataset of their choosing (e.g. privacy sensitive data, such as a corporate filesystem, or personal email).
3. Users or organisations can upload (or “push”) a custom index or custom metadata to contribute to the OWI.

The core data structure behind the OWI is the inverted file, a mapping between each term on the Web and the pages or documents in which it appears. To ensure usability of the OWI, we have chosen to store the inverted files in a common, easily transferable format: the Common Index File Format (CIFF)[1].

CIFF is a Protobuf schema that describes the inverted files in a structured, consistent and minimal format. A CIFF file consists of the following data:

- A header, containing basic statistics about the collection (e.g. the number of documents, the number of unique terms, the average document length, etc.).
- For each term in the corpus, a record with the document frequency (how many documents contain the term), the collection frequency (how often the term occurs in total throughout the collection), and a posting list. The posting list has an entry for each document in which the term appears, containing an internal document identifier and the term frequency (how often the term occurs in that document).
- For each document in the corpus, a record with the internal, numeric document identifier, the external document identifier (allowing the document to be mapped back to the original corpus), and the document length.

The CIFF standard contains the basic information needed to build a successful search engine using an index, which makes it easy for an existing search engine to import the data and transform it to their internal data structures — easier than transforming it from one search engine’s internal format into another’s. In fact, the CIFF standard was proposed by the developers of a number of existing open source search engines (like Lucene[2], (Py)Terrier[3], and PISA[4]), and these search engines already support reading and/or writing to CIFF. As a result, the indexes we build in the project can readily be used by external parties with minimal extra effort.

As part of our work on and with the OWI, we also investigate limitations of the CIFF standard, and propose extensions of the format where necessary.

[1] Lin, Jimmy, et al. "Supporting Interoperability Between Open-Source Search Engines with the Common Index File Format". *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020.

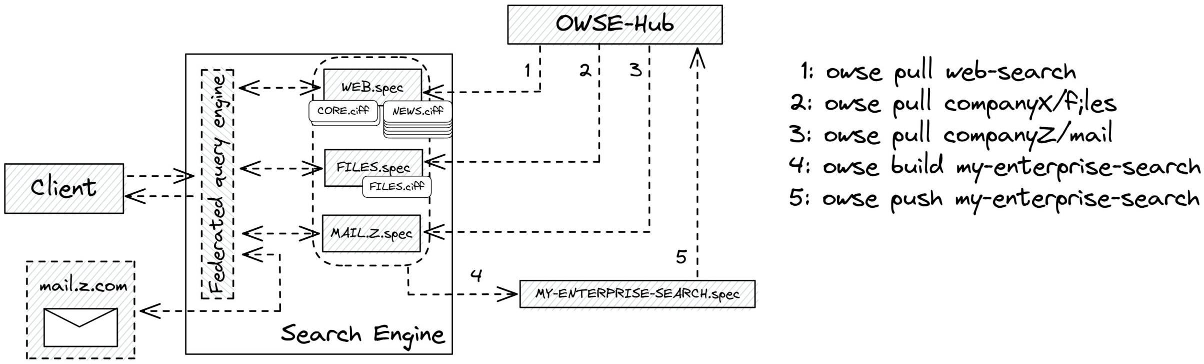
[2] Apache Software Foundation (2000). Lucene [Computer software]. Retrieved from https://lucene.apache.org/

[3] Ounis, Iadh, et al. "Terrier: A High Performance and Scalable Information Retrieval Platform". *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*. 2006.

[4] Mallia, Antonio, et al. "PISA: Performant Indexes and Search for Academia". *Proceedings of the Open-Source IR Replicability Challenge co-located with 42nd International {ACM} {SIGIR} Conference on Research and Development in Information Retrieval, OSIRRC@SIGIR 2019, Paris, France, July 25, 2019*. 2019.

1.2. Vision for the OpenWebSearch Engine Hub

To define downstream search engines using the OWI, we will also introduce the Open Web Search Engine Hub (OWSE-HUB). Similar to the OWI, the OWSE-HUB forms a web-based information system comparable to the Docker hub, but it will contain complete search engine stacks to enable the fast and easy creation of new search verticals. The architecture of the envisioned OWSE-Hub is outlined in Figure 2.



```
 1: owse pull web-search
2: owse pull companyX/files
3: owse pull companyZ/mail
4: owse build my-enterprise-search
5: owse push my-enterprise-search
```

Figure 2: General architecture of the OWSE-Hub, and how the (federated) search strategy would capture declaratively the usage of indexes retrieved through specifications in the OWSE-Hub. Users can (1-3) pull search engine stacks, (4) build their own specifications for a (composite) search engine, and (5) push specifications to share with others.

Using the OWSE-Hub, users can declaratively define their own search configurations. Users can “pull” pre-defined specifications from the OWSE-Hub, use those to “build” their own custom search engines, and “push” the most useful ones to share these with others. This flexible setup allows for the creation of a wide variety of search engines, not only for commercial usage but also for personal and corporate search, and allowing both centralized and federated search setups.

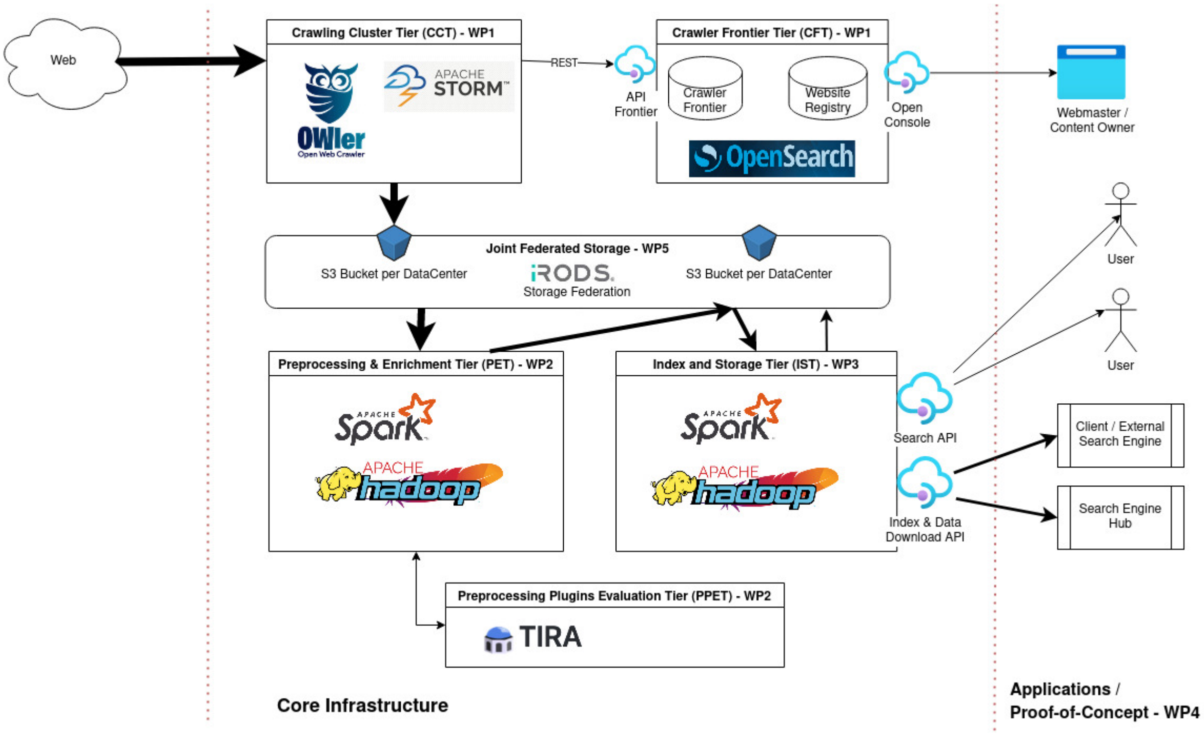
The way in which the search specifications are to be defined, how they can be used and deployed, and the format in which they can be exchanged, has not yet concretely been decided. We will start working on this in the second year of the project, when the actual work on the OWSE-HUB is set to start.

2. The Open Web Index

For year one, our goal for the OWI was to develop and deploy the first full pipeline that crawls (a part of) the Web, pre-processes and indexes the web content, and makes the resulting indexes available in shared storage.

2.1 Overview

We define the architecture for the full pipeline in terms of *cluster tiers*. A cluster tier is defined as *a well-defined set of services* (and corresponding software-stack) that *must* run in the same cluster / set of virtual machines. Figure 3 outlines a first version for cluster tiers and their interplay.



Web
Crawling Cluster Tier (CCT) - WP1
OWler
Open Web Crawler
APACHE STORM
REST
API Frontier
Crawler Frontier Tier (CFT) - WP1
Crawler Frontier
Website Registry
OpenSearch
Open Console
Webmaster / Content Owner
Joint Federated Storage - WP5
S3 Bucket per DataCenter
iRODS
Storage Federation
S3 Bucket per DataCenter
User
User
Preprocessing & Enrichment Tier (PET) - WP2
Spark
APACHE hadoop
Index and Storage Tier (IST) - WP3
Spark
APACHE hadoop
Search API
Index & Data Download API
Client / External Search Engine
Search Engine Hub
Preprocessing Plugins Evaluation Tier (PPET) - WP2
TIRA
Core Infrastructure
Applications / Proof-of-Concept - WP4

Figure 3: Cluster tiers in the core architecture behind the OWI. A cluster tier is defined as a set of machines running a well-defined software stack in the same physical data-center for a particular purpose with clear, well defined interfaces to other cluster tiers

We describe the different cluster tiers in more detail below.

2.2 Crawling (WP 1)

Crawling takes place on the following two cluster tiers:

- The Crawling Cluster Tier (CCT) contains a crawler responsible for crawling the Web, thereby creating WARC files (i.e., collections of HTTP streams from the crawling process). These WARC[5] files are stored in an S3 bucket[6] that can be accessed by the other tiers. There can be multiple CCT, but there should be only one CCT per data center, as the CCT can scale with the amount of resources provided.
- The Crawler Frontier Tier (CFT) is a singleton tier (i.e., only one instance exists across all participating data centers). It coordinates the multi-tier crawling process, keeping track of crawled URLs, access statistics and cache digest. The crawler frontier is the primary data source for the website registry, which provides the interface for interaction between webmasters and the frontier (e.g., an interface for take-down requests).

[5] International Organization for Standardization. (2017). *Information and documentation — WARC file format* (ISO/DIS Standard No. 28500). Retrieved from https://www.iso.org/standard/68004.html

[6] Amazon AWS S3. Available: https://aws.amazon.com/s3/

The CCT consists of the OWLer (OWS.eu crawler), a distributed web crawler built with StormCrawler[7] running on Apache Storm, at each individual data center. Aside from regular crawling functionality, the OWLer also includes a classifier for distinguishing between benign, malicious and adult content.

The CFT consists of a single OpenSearch[8] instance running at one of the data centers. It also contains a metrics server which aggregates logs from different stages and allows for (limited) statistical analysis.

Currently (end of July, 2023), the CCT has been deployed successfully at three partner sites (UNI PASSAU, BADW-LRZ and IT4I@VSB) and is in operation at two (UNI PASSAU and BADW-LRZ). The CFT is deployed at CERN. At the time of writing, the CCT crawls roughly 200GB of content per day but can be scaled up by using more machines and we aim to scale it up to 1 TB/day until end of year. This requires also to resolve security related issues, like raising false positive network security alarms when accessing botnets. However, the frontier still remains a bottleneck to be scaled up accordingly with the number of machines used.

2.3 Preprocessing and content analysis (WP 2)

Preprocessing and content analysis encompasses the following two cluster tiers:

- The Preprocessing and Enrichment Tier (PET) takes the WARC files from the S3 Bucket filled by the CCT and extracts cleaned HTML and metadata. After the first year, the metadata consists of the page's language and several properties derived from the URL. Eventually, more metadata — like topic, geo-information and entities — will also be extracted. Following the partitioning of the CCT, each data center should have a dedicated PET for the WARC files stored at that data center. The metadata extracted by the PET is stored in Parquet format.
- The Preprocessing Plugins Evaluation Tier (PPET) is another singleton tier that enables the evaluation of plugins for the content analysis library. In order to expand enrichment capabilities, both project members as well as third parties may develop plugins to be used in the PET.

  To ensure good quality and sufficient throughput, candidate plugins first have to perform their enrichment task on problem-specific benchmarking data (e.g. a classification dataset for a plugin performing webpage classification), using an instance of the TIRA[9] platform hosted at one of the data centers. This platform provides the means for evaluation as a service with a focus on information retrieval research. It can host shared tasks on a given research problem, run submitted software in virtual machines and thereby create reproducible results.

The PET uses Apache Spark batch jobs that apply Resiliparse[10] to parse and clean the HTML content and apply several types of metadata enrichment. The work on this tier is ongoing and will be outlined in further detail in deliverable D2.1.

[7] DigitalPebble, Ltd. (2014). StormCrawler [Computer software]. Retrieved from https://stormcrawler.net

[8] Amazon Web Services (2021). OpenSearch [Computer software]. Retrieved from https://opensearch.org/

[9] Fröbe, Maik, et al. "Continuous integration for reproducible shared tasks with TIRA. io." *European Conference on Information Retrieval*. Cham: Springer Nature Switzerland, 2023.

[10] Bevendorff, Janek, et al. "Elastic chatnoir: Search engine for the clueweb and the common crawl." *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*. Springer International Publishing, 2018.

The PET runs currently (end of July, 2023) at the same infrastructure sites as the CCT (BADW-LRZ and IT4I@VSB). Functionality is in place to estimate the amount of resources required for processing a day's worth of WARC files, such that the PET can scale along with the amount of content being crawled by the CCT.

The PPET builds upon TIRA, a platform for replicability and comparison of information retrieval experiments. Currently, TIRA is hosted at WEBIS and is being used for the evaluation of several shared tasks at well-known IR conferences. We are looking into hosting a TIRA instance at one of the infrastructure partners, to serve as the core part of the PPET to evaluate PET plugins.

2.4 Indexing (WP 3)

Indexing is supported in the Index and Storage Tier (IST).

The IST turns the cleaned content from the PET into a usable index/inverted file. Similar to the CCT and the PET, each data center should have its own IST (and only one), responsible for creating an index for the documents crawled by that data center. Indexes are inverted files partitioned into so-called "shards" by selected types of metadata derived by the PET (e.g., topic and language), distributed as CIFF[11] files.

Similar to the PET, the IST is implemented as a Spark batch job. In its current state, it reads the Parquet files delivered by the PET and writes out inverted files per shard (by arbitrary metadata values). This allows us to build semantically coherent shards of the full web index (dependent on the metadata extracted by the PET), which can be used to enable a large variety of downstream search engines. For instance, language can be used to build search engines for specific countries, geo-information can be used to focus on specific areas, and classified topics or genres can be used to build search engines focusing on a specific area (like news and/or sports).

Currently (end of July, 2023) the IST has successfully been deployed to both BADW-LRZ and IT4I@VSB. Current efforts are focusing on scaling up the indexer and building indexes for the content crawled thus far.

2.5 Infrastructure (WP 5)

At the time of writing (end of July, 2023), the full technology stack behind the OWI has been deployed at BADW-LRZ and IT4I@VSB and is planned to be extended to CSC's stack. The CFT and corresponding metrics services run at CERN. The different services are currently secured via IP address and thus not accessible from outside the participating services / VPN connections to the computing centers.

A core contribution to this successful deployment is the development of a consistent data infrastructure across all data centers. The data infrastructure is built with iRODS[12], allowing for seamless federated access to the data at each of the data centers. On top of the iRODS layer, each data center runs Minio[13] to provide a uniform S3-like access point to the OWS data.

[11] Lin, Jimmy, et al. "Supporting Interoperability Between Open-Source Search Engines with the Common Index File Format". *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020.

[12] iRODS Consortium. iRODS [Computer software]. Retrieved from https://irods.org/

[13] MinIO, Inc (2016). MinIO [Computer software]. Retrieved from https://min.io/

The CCT, PET and IST share relevant data between each other by writing to and reading from S3 buckets in this federated data architecture. To ensure consistency and prevent unnecessary data transfer between data centers, we have agreed upon a shared directory structure adhering to the following principles:

- Every data center, or zone, has a dedicated top level directory containing data that is physically located at that location. This means the CCT, PET and IST can read from and write to these locations for local data processing.
- There exists a directory for the crawled WARC files, one for the cleaned Parquet files, and one for the indexed CIFF files. Each of these contains a subdirectory for every day, of the form YYYY-MM/DD, representing the documents crawled/cleaned/indexed on that specific day.
- For data that requires replication across data centers, we use a “shared” top level directory.
- For data that requires sharing outside the consortium, we use a “public” top level directory.
- For data contributed by the community, we use a “community” top level directory.

2.6 The actual Open Web Index

The final product of the full pipeline described above — the Open Web Index — consists of the CIFF files in the zone-specific S3 buckets (see also Figure 1). Access control to the OWI will be arranged through B2ACCESS[14].

To use the Open Web Index, downstream search engines will be able to download the CIFF files they are interested in (e.g. with a specific language or from a specific date range), and import them into a search engine of their choice. Alongside the CIFF files, the metadata Parquet files can be downloaded for additional use in a search engine. For instance, the cleaned text can be used for snippet extraction, and the metadata fields can be used to enrich or filter the search results obtained by a full-text search on the index.

3. The OpenWebSearch Engine Hub

In the first year of the project, the main focus has been put on developing and deploying a first version of the Open Web Index. As a result, the OWSE-HUB has not yet been developed. This is in line with the schedule of the project: we start working on the OWSE-HUB in M12.

Nevertheless, first steps towards building the OWSE-HUB have been taken, by integrating the indexer (the IST) and the CIFF standard into TIRA. This has given us a first glimpse into how we can use the indexes constructed by the IST in a centralized platform, helpful to make our vision for the OWSE-HUB more concrete.

4. Outputs and Statistics

4.1. Overview of open-source software outputs

Each of the parts we developed to build and sustain the OWI and OWSE-HUB is published as an open-source software project. Table 1 lists all of these projects and includes links to their respective Git repositories, as well as a version archived on Zenodo near the end of the first year.

[14] B2ACCESS. Available: https://www.eudat.eu/catalogue/b2access

Table 1: Open-source software packages developed to build and use the OWI and OWSE-HUB.

Title	Open Web Search Crawler (OWler) (see also D1.1)
Repository	https://opencode.it4i.eu/openwebsearcheu-public/owler
Archived	https://doi.org/10.5281/zenodo.8278932

Title	Resiliparse: WARC/HTML parsing and preprocessing library
Repository	https://github.com/chatnoir-eu/chatnoir-resiliparse
Archived	https://doi.org/10.5281/zenodo.8262470

Title	Preprocessing at scale: Resiliparse in large Spark batch jobs
Repository	https://opencode.it4i.eu/openwebsearcheu-public/preprocessing-pipeline
Archived	https://doi.org/10.5281/zenodo.8262403

Title	TIRA: platform for replicability and comparison of information retrieval experiments
Repository	https://github.com/tira-io/tira
Archived	https://doi.org/10.5281/zenodo.8261862

Title	Open Web Search Indexer
Repository	https://opencode.it4i.eu/openwebsearcheu-public/spark-indexer
Archived	https://doi.org/10.5281/zenodo.8261098

Title	CIFF toolkit: library for processing WARC files
Repository	https://opencode.it4i.eu/openwebsearcheu-public/ciff-toolkit/
Archived	https://doi.org/10.5281/zenodo.8261147

Title	CIFF to Lucene index converter

Repository	https://github.com/informagi/lucene-ciff
Archived	https://doi.org/10.5281/zenodo.8261333

4.2 Statistics of the Open Web Index

Table 2 shows some statistics of the OWI, measured at the end of July. These help give an overview of the amount of data we have been able to process towards the end of the first year.

Table 2: Statistics of the OWI, as measured on July 31st 2023.

Documents crawled	20M-40M / day
Documents preprocessed	~ 475M
Documents indexed	~ 300M
Number of unique languages	179
Number of unique domains	~ 10M

4.3 Top domains crawled

Table 3 shows the top domains crawled in the time between 12th July 2023 and 24th July 2023.

Table 3: Top domains crawled in the time between 12th July 2023 and 24th July 2023.

Domain	Fraction	Absolute Count [for 12 days]
blogspot.com	6.06%	761425
wordpress.com	4.22%	530621
wikipedia.org	2.64%	331622
europa.eu	1.11%	139832
yahoo.com	0.93%	116319
hatenablog.com	0.89%	111458
mit.edu	0.62%	78211
nih.gov	0.60%	75568
web.app	0.58%	73432
free.fr	0.57%	71916
google.com	0.55%	69052
fc2.com	0.51%	64462
harvard.edu	0.51%	64019
microsoft.com	0.45%	56230
altervista.org	0.45%	56224
uk.com	0.44%	55700
nasa.gov	0.43%	54270
libsyn.com	0.43%	53579
stanford.edu	0.42%	52404
home.blog	0.42%	52362
pinterest.com	0.41%	52016

amazonaws.com	0.41%	51201
appspot.com	0.39%	48906
berkeley.edu	0.37%	47045
mpg.de	0.37%	46842
airbnb.com	0.36%	45529
ox.ac.uk	0.33%	41425
sakura.ne.jp	0.32%	40730
alibaba.com	0.32%	39906
typepad.com	0.32%	39587
wordpress.org	0.31%	39448
wiktionary.org	0.31%	39058
nsw.gov.au	0.31%	38565
aif.ru	0.31%	38481
medium.com	0.30%	38251
cam.ac.uk	0.30%	37736
cornell.edu	0.30%	37284
googlesource.com	0.30%	37282
umich.edu	0.30%	37257
archive.org	0.29%	37020
debian.org	0.29%	36675
workplace.com	0.28%	34949
apple.com	0.28%	34754
us.com	0.27%	34526
wikidot.com	0.26%	33165

5. Conclusion and Outlook

In this document, we have summarized our work on setting up the Open Web Index in the first year of project OpenWebSearch.EU. We have managed to build a full pipeline — from crawling the Web to preprocessing and indexing — that has been deployed at more than one infrastructure partner. At the end of July 2023, roughly 300 million webpages have successfully been indexed, partitioned across 179 languages.

For year 2 of the project, we will focus on improving and expanding the current pipeline. For instance, we will extend deployment of the full pipeline to the other infrastructure partners in the consortium, to make optimal use of the large amount of resources we have available to us. Also, we will measure performance and throughput of the current cluster tiers, to evaluate how well they scale and determine where further improvements are necessary.

We briefly discuss specific plans for each of the cluster tiers below.

5.1 Crawling

In year two (Y2), we're planning some improvements for our web crawler system. First, we want to boost the Crawling Cluster Tier (CCT) by adding more machines, so it can gather between 5 to 10 TB of data daily. Next, we'll make it easier for other crawlers to connect with our URL Frontier, allowing them to share their web data in WARC files. We also plan to add new features to our crawler, so it can focus on specific topics when collecting data. Lastly, we'll work on improving the speed and efficiency of the URL Frontier to ensure everything runs smoothly and scales up easily.

Beyond the development of the crawler our focus will be on legally compliant crawling. This includes the development of the open webmaster console as well as defining metadata standards for expressing privacy and copyright issues and filling those from terms of us and privacy statements of websites using machine learning.

5.2 Preprocessing and content analysis

A central goal for year two is to ensure the scalability of the PET in response to the number of WARC files crawled by the CET. In addition to that, we will conduct research on semantic enrichment of web content under the umbrella term “computational ethics” that deals with the detection of disruptive or potentially harmful content.
The goal for the PPET is to have a running instance on one of the data centers to benchmark and review potential plugins to the PET.

5.3 Indexing

Semantic sharding of the OWI should enable a wide variety of downstream search applications. We will research how we can efficiently and effectively construct these shards, and measure their coherence. To prevent downstream search engines from having to download the full index every once in a while, we will also investigate how to properly handle index updates and duplicate pages, as well as how to indicate that pages should be deleted from the index.

We will continue to work closely with WP4 (Search Applications and User Experience) to investigate how the OWI can be used to enable different search applications and paradigms.

5.4 OWSE-HUB

In the coming year, we will develop more concrete plans for the OWSE-HUB. Specifically, we will define how search engine specifications are constructed and represented, and how we can set up the platform (or ‘hub’) for sharing these specifications.

Our target is to deploy an initial version of the OWSE-HUB by the end of year 2.