Open WebSearch

OWLer: Preliminary results for building a collaborative Open Web Crawler

Open Web Search

Funded by the European Union

SUPPORTED BY NGI

Who are we?

OpenWebSearch Team @ Uni Passau

Michael Granitzer

Jelena Mitrović

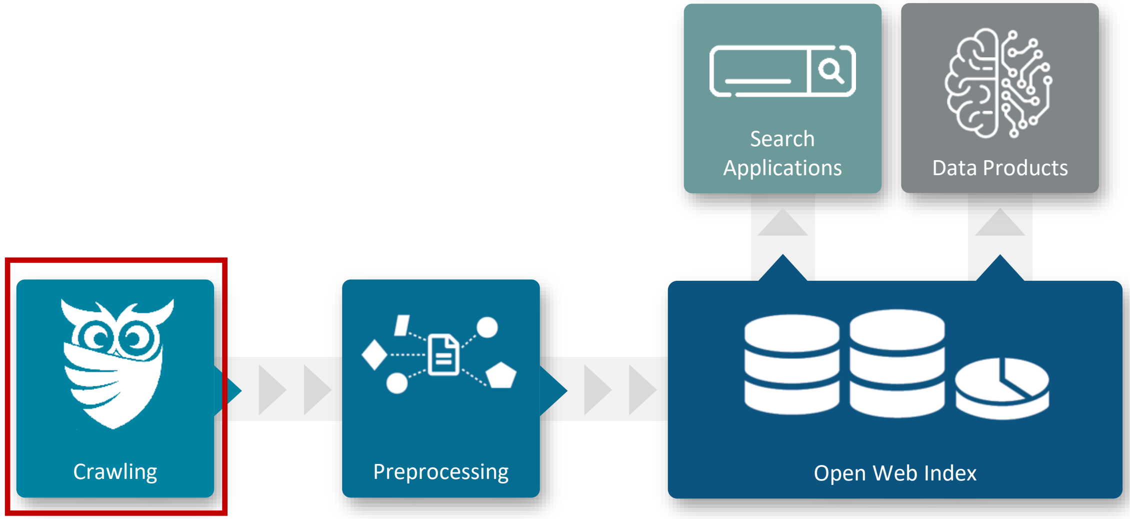
Saber Zerhoudi

Michael Dinzinger

„A Comprehensive Dataset for Webpage Classification“
Mohammed Al-Maamari, Mahmoud Istaiti

ows.eu/owler

OpenWebSearch.eu
Collaboratively building an Open Web Index (OWI)



➜ **What were the prerequisites for our work?**

➜ Project principles: **Openness**, **Transparency** and **Collaboration**

- Using and supporting Open Source software projects

➜ Heterogeneous infrastructure distributed over serveral European datacenters.

➜ High expectations on ourselves regarding the size of the final Web Index

Open Web Crawler (OWLer)
An incremental and distributed web crawling system

OWLer is a derivative of StormCrawler. It documents its fetch activities as WARC files and feeds the OWI indexing pipeline.

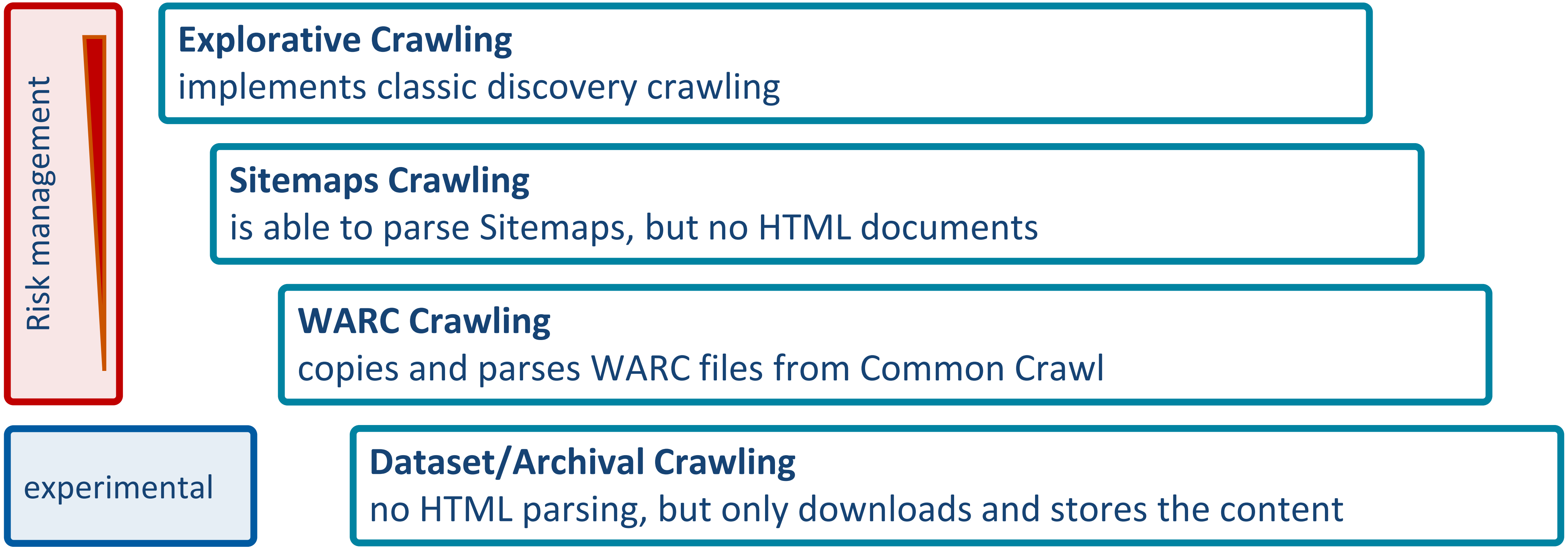
StormCrawler is a popular and mature Open Source web crawler, written in Java.

Apache Storm serves as a distribution layer below StormCrawler, making it lightweight and scalable.

OWLer StormCrawler

➜ StormCrawler allows to obtain **high performance** despite the use of commodity hardware thanks to Apache Storm.

➜ The crawling pipeline is **highly customizable**, which helps us to meet the broad spectrum of requirements, ranging from general-purpose discovery crawling to task-specific dataset crawling.



Risk management

Explorative Crawling
implements classic discovery crawling

Sitemaps Crawling
is able to parse Sitemaps, but no HTML documents

WARC Crawling
copies and parses WARC files from Common Crawl

experimental

Dataset/Archival Crawling
no HTML parsing, but only downloads and stores the content

Open Web Crawler (OWLer)
An incremental and distributed web crawling system

URL Frontier communicates with all crawlers, and queues the URLs for the next fetch.

OWLer is a derivative of StormCrawler. It documents its fetch activities as WARC files and feeds the OWI indexing pipeline.

StormCrawler is a popular and mature Open Source web crawler, written in Java.

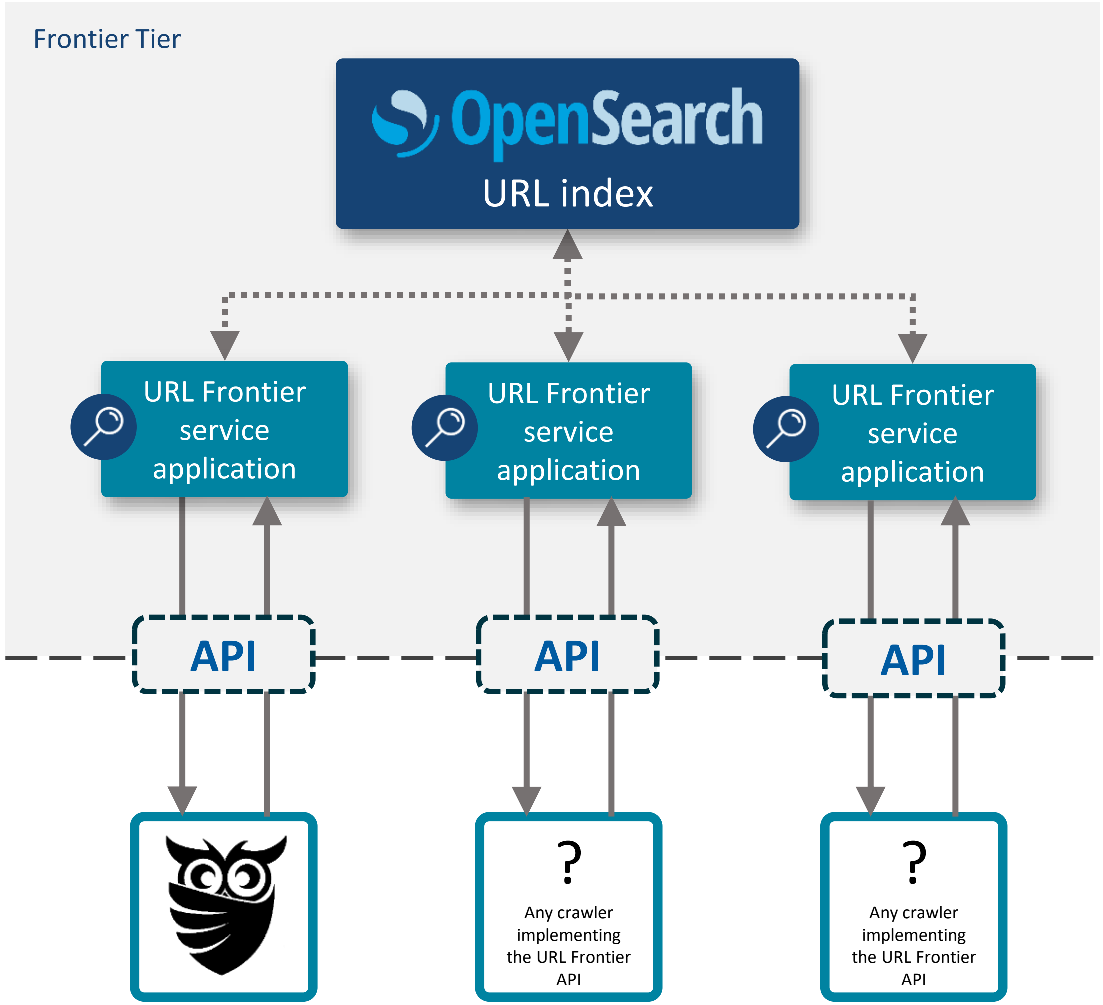
Apache Storm serves as a distribution layer below StormCrawler, making it lightweight and scalable.

OWLer URL Frontier

- ➜ The URL Frontier framework **stores and queues the crawl URLs** in an incremental crawling system.
- ➜ The URL Frontier ensures a crawl delay per domain (**politeness**) and revises the order of outgoing URLs (**priorization**).
- ➜ It comprises:
  - – An OpenSearch index for storing URLs, and
  - – Multiple service applications for communicating with the crawlers

Shortcoming in joint peer-to-peer crawling:
The Frontier services partition the crawl space equally among them with the help of a simple consistent hashing algorithm.

- ➜ Enabling Frontier services to specify a scope of interest

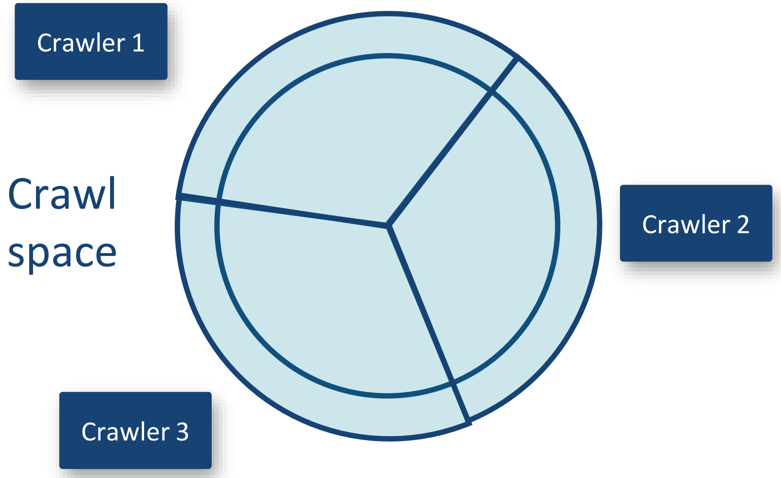

Frontier Tier
OpenSearch
URL index
URL Frontier service application
URL Frontier service application
URL Frontier service application
API
API
API
?
Any crawler implementing the URL Frontier API
?
Any crawler implementing the URL Frontier API

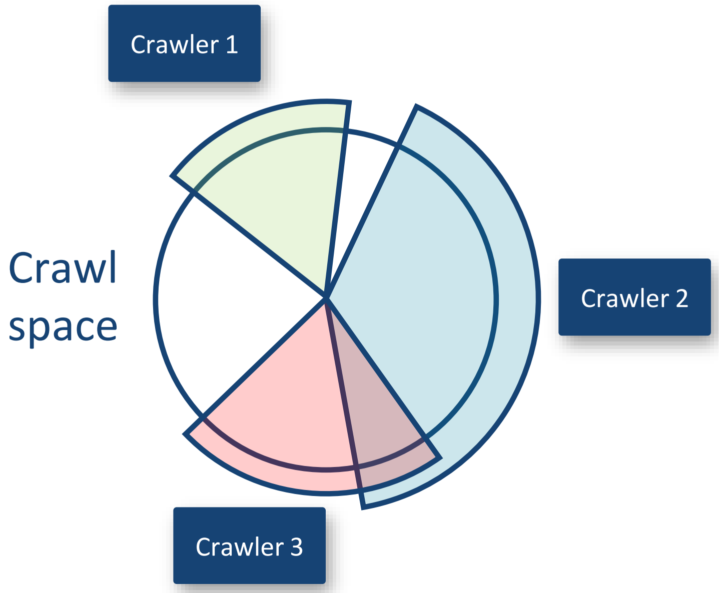
Towards more comprehensive collaborative crawling
Conclusion

➜ The specification of a scope of interest for Frontier service facilitates:

- Risk management („Give me only URLs, which have been discovered through Sitemaps“)
- Blacklisting of URLs („Give me only URLs, which have not been marked as Blacklisted“)
- The creation of topic-specific indices („Give me only URLs for physics-related content with TLD .ch“)
- More fine-grained partitioning of crawl space for better utilizing the heterogeneous infrastructure

➜ The modifications allow for a large-scale peer-to-peer crawling with multiple agents collaborating in the same shared crawl





Some numbers

- ➜ We started crawling about 4-5 months ago, first with one experimental node scaling up to at most **five crawlers** at the same time.
- ➜ In the last four weeks, a single crawler has fetched around **120M web documents**, 93% of them successful fetches (HTTP status code 200).
- ➜ In August and September, we have produced **over 20 TiB of WARC files** for feeding the OWI pipeline. During this time, we have copied around **780M WARC entries from Common Crawl**.
- ➜ The Frontier index is filled with **over 3.5 billion URLs**, with up to 24 metadata fields per URL.

Upcoming work

- ➜ Implementing a Crawling-On-Demand functionality similar to IndexNow
- ➜ Research on legally compliant and license-aware crawling
- ➜ Increasing the Index quality through near-duplicate elimination of URLs and more advanced priorization of URLs

Questions?

Open Source Software Repositories:

StormCrawler

Opensearch implementation of URLFrontier service