

The Archived Web as Research Data, The BelgicaWeb Story



Presentation outline

1. Introduction & objectifs
2. Technical stack and data capture process
3. Server infrastructure
4. Data enrichment
5. Legal aspects
6. Next steps
7. Q&A

INTRODUCTION & OBJECTIFS



THE BELGICAWEB PROJECT



- 2-year research project funded by BRAIN-be (BELSPO)
- **Objectives:**
 - Investigate how to sustainably provide **access** to these collections
 - Develop the necessary data infrastructure (**API & access platform**)
 - **Enrich** the (meta)data
 - Analyse the relevant **legal frameworks**
 - **Promote** Belgium's born-digital heritage
- Collaboration with a panel of experts

TECHNICAL STACK & DATA CAPTURE PROCESS



Technical stack

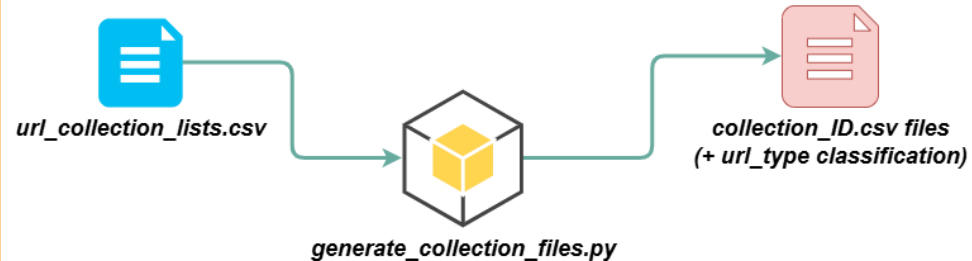
Regarding the architecture, BelgicaWeb is built on:

- Open-source **Symfony** framework for the platform
 - Symfony's companion framework for APIs, **API Platform**
 - All in **PHP 8.4**
 - The database is **MariaDB**
 - Use of **Stimulus** for UX and JavaScript component integration
-
- For indexing and replay, **SolrWayback** 5.4.1
 - To capture web content **Browsertrix Crawler**

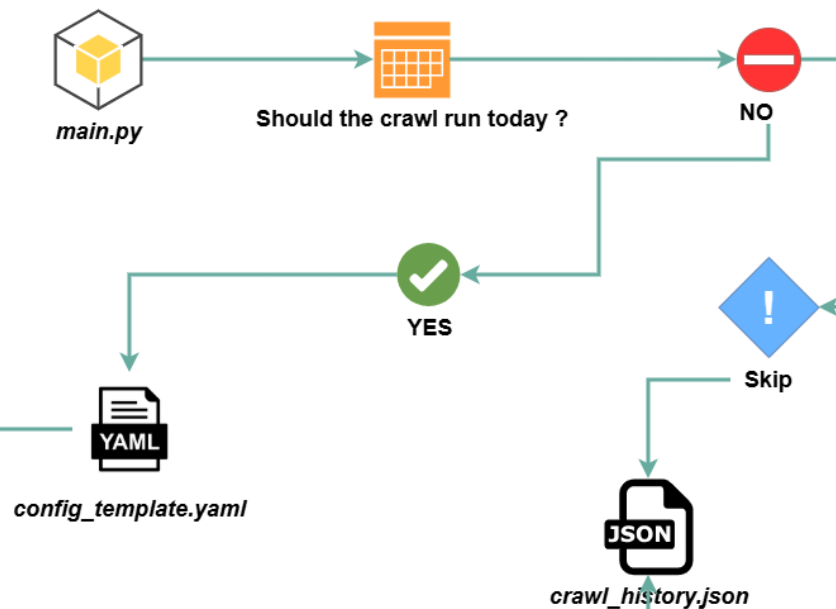
Data capture and process

- **Scope Expansion (2026):** now reaching 27 collections and ~ 2,000 seeds/url.
- **Continuous Data Capture:** operations are still active.
- **Technology Strategy:**
 - Standardized on Browsertrix-crawler for all content types.

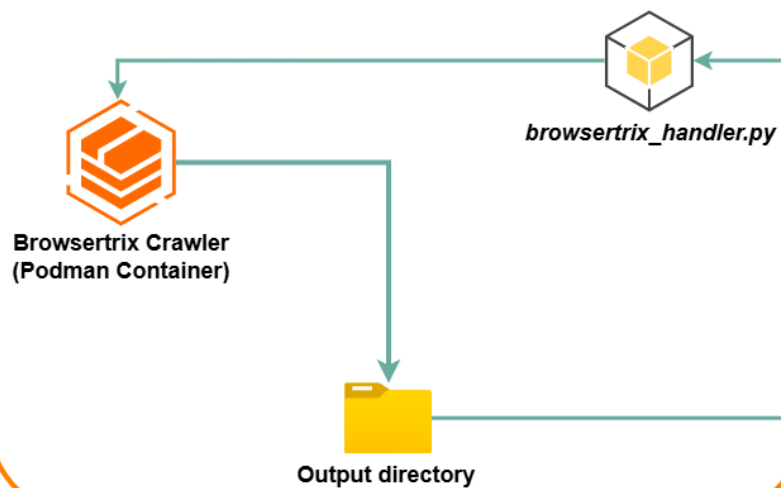
Phase 1 - Data Preparation



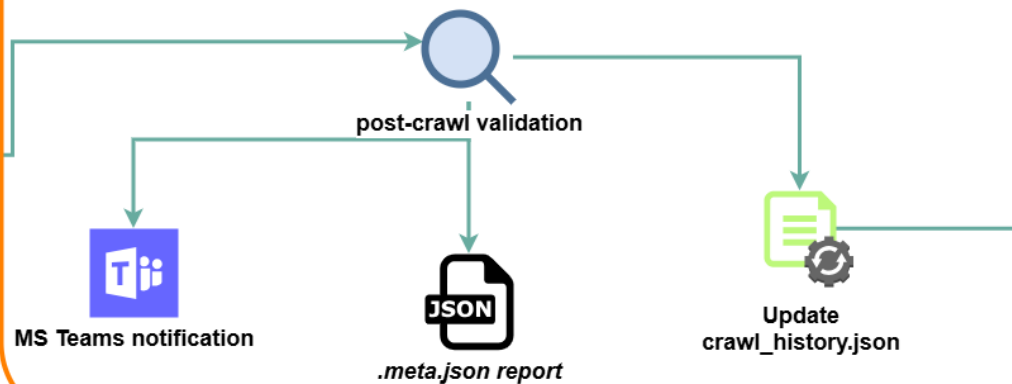
Phase 2 - Orchestration & scheduling



Phase 3 - Crawl Execution



Phase 4 - Validation & Reporting



SERVER INFRASTRUCTURE

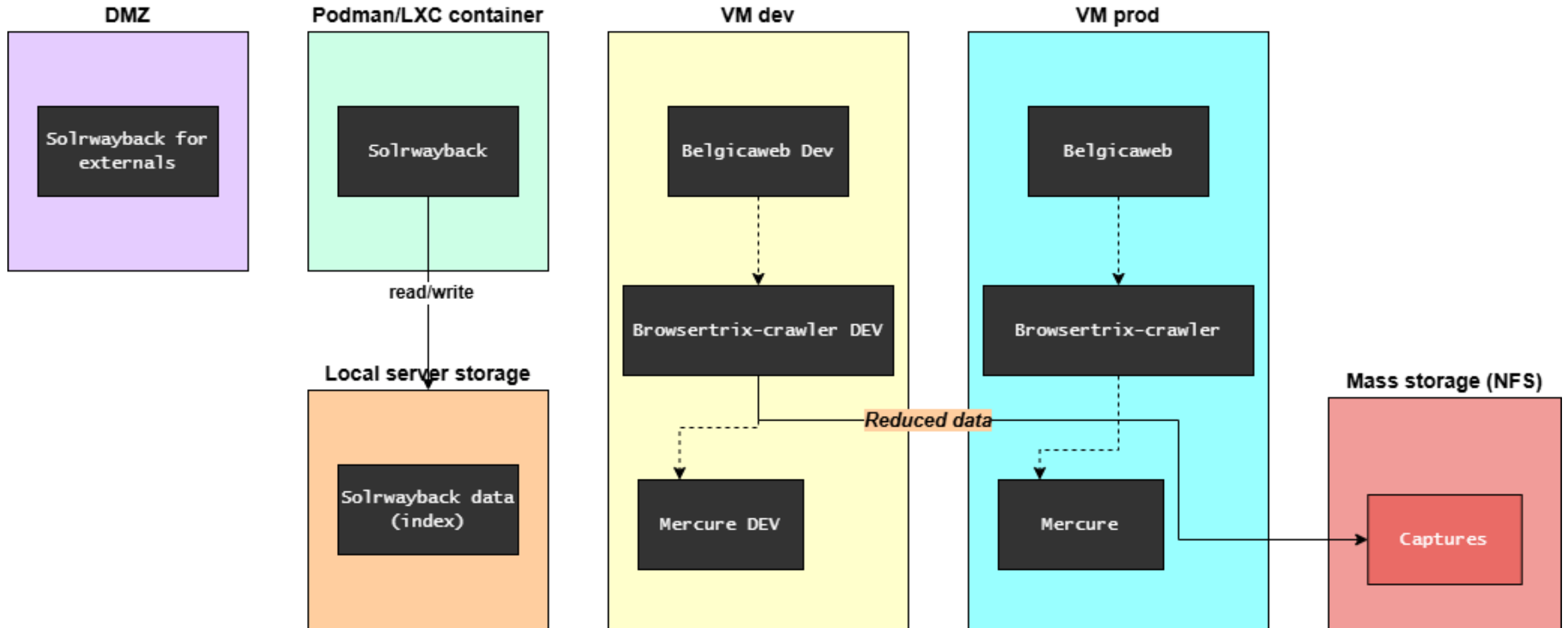


For the geeks, our server configuration...

2 x PowerEdge R760 Infrastructure with this specifications:

- **Hardware:**
 - **Model:** Dell PowerEdge R760
 - **RAM:** 64 GB
 - **Network:** 2x 4-port 10GbE NICs
- **Storage configuration:**
 - **OS storage (proxmox 9.1):**
 - ✓ 2 x 447 GB SATA SSD
 - ✓ Configuration: ZFS RAID1 (mirror)
 - **Data storage (VMs & containers)**
 - ✓ 2 x 2.2 TB drives
 - ✓ Configuration: ZFS RAID1 (mirror)

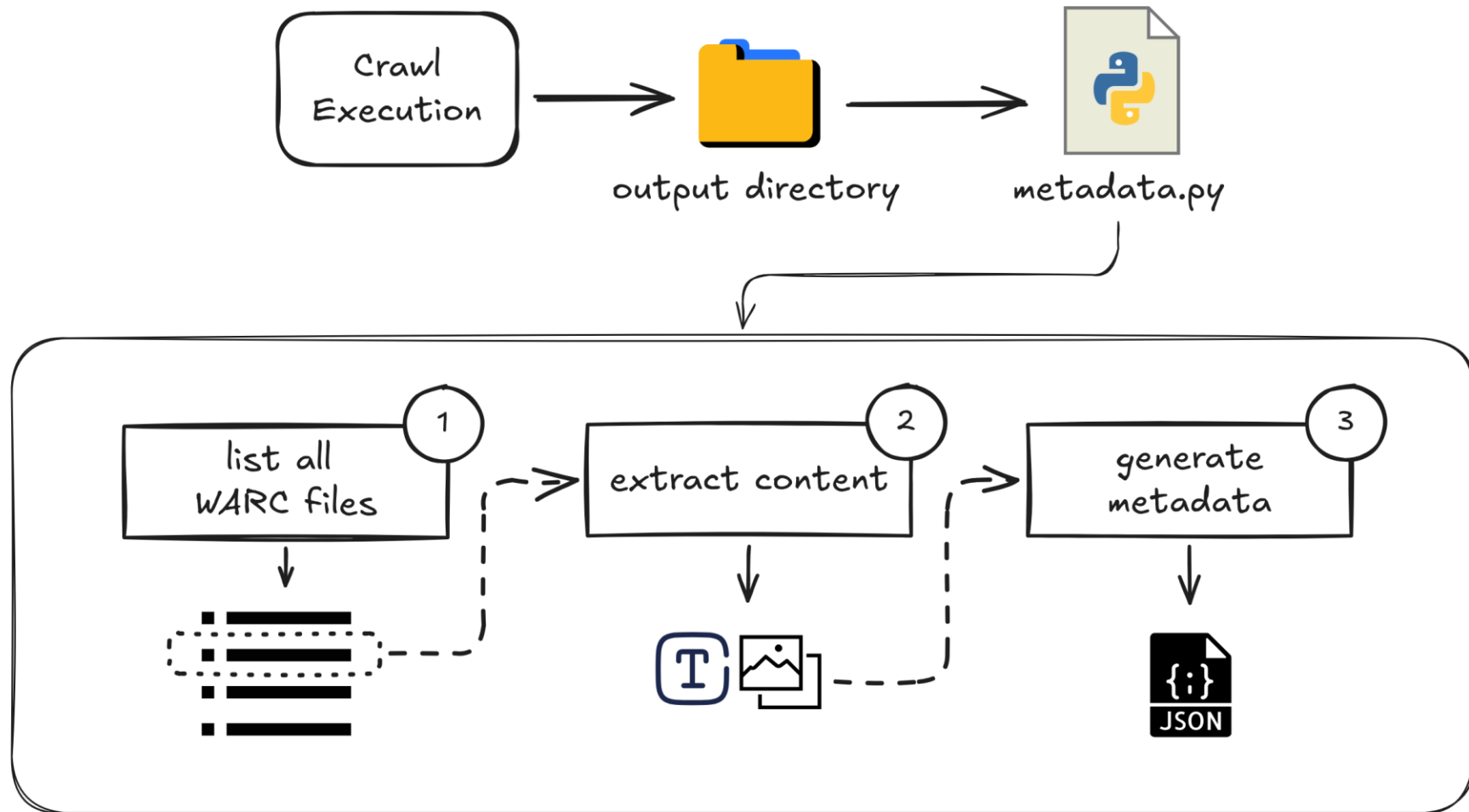
... and the architecture in one schema !



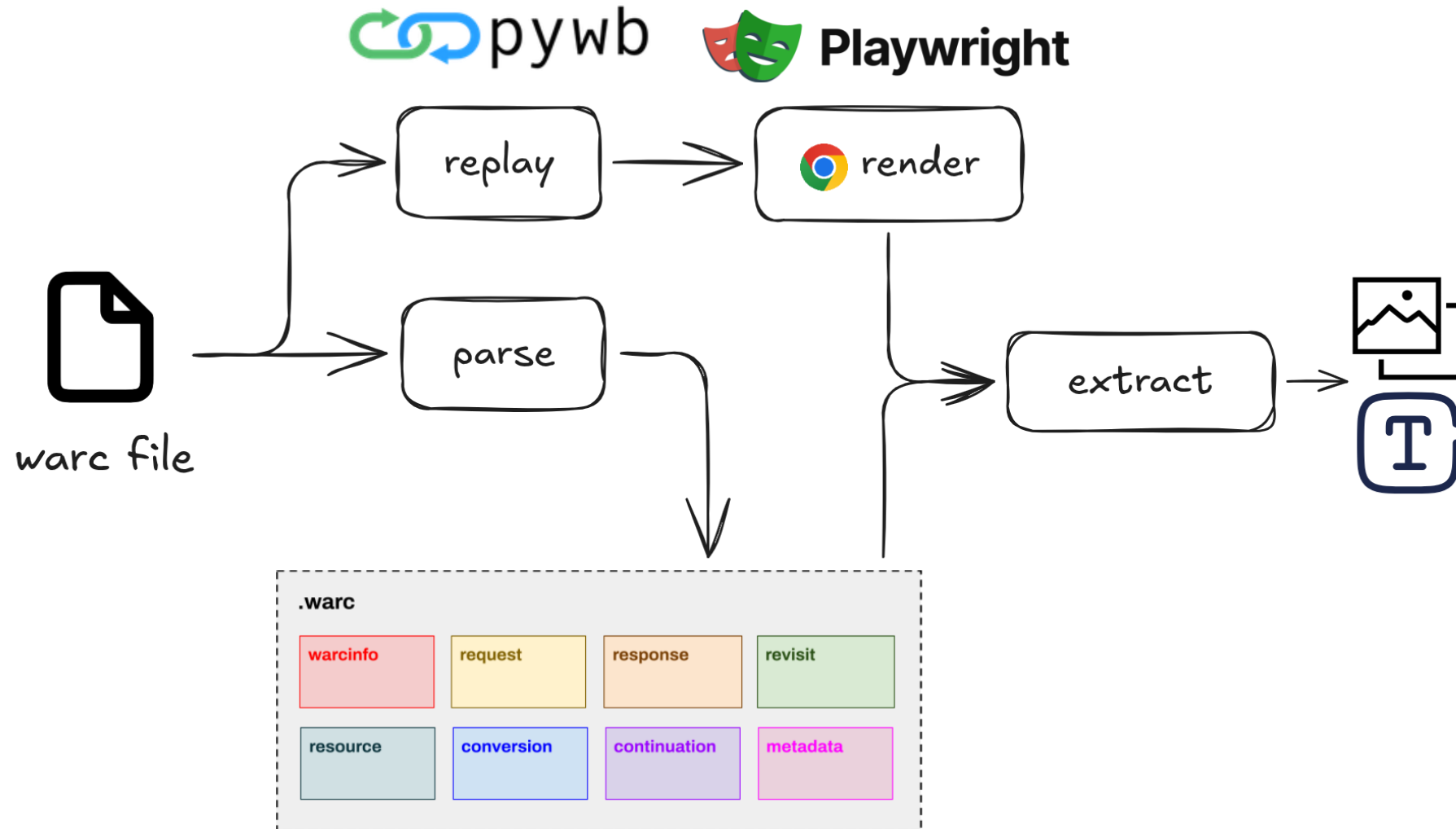
DATA ENRICHMENT



Data enrichment workflow



Data enrichment workflow



KEY FEATURES



Born-digital heritage is digitally created and has no analogue equivalent. Rapidly changing technology makes this type of heritage especially vulnerable. BelgicaWeb is a [BRAIN 2.0 project](#) funded by BELSPO. BelgicaWeb aims to make Belgium's born-digital heritage accessible and FAIR, i.e. **F**indable, **A**ccessible, **I**nteroperable and **R**eusable. We want to do this by developing a user-friendly access platform and an API that enables access at data level.

- 1. investigate how to provide **sustainable access** to born-digital collections,
- 2. develop the necessary **data infrastructure** to access such collections,
- 3. **enrich the (meta-)data**,
- 4. analyse the relevant **legal frameworks** and
- 5. **promote** Belgium's born-digital heritage.

Project phases and duration

The BelgicaWeb project started in April 2024 and is scheduled to end in April 2026. During these two years, we will work on among other things:

- 1. Creating **collections of born-digital data** using tools for archiving web and social media content.
- 2. Developing the **API and access platform** in two phases: a pilot and a final phase.
- 3. **Indexing and enriching data** by using digital methods such as Linked Data and Natural Language Processing.
- 4. An analysis of the **legal frameworks** related to data sharing, text and data mining, data protection and privacy rights, freedom of expression and the impact of the proposed European regulation on artificial intelligence.



Reference group of experts

Within the project, a reference group of experts will provide iterative input on the selection, development of the API and access platform, data enrichment and quality control and usability.

Project partners

The BelgicaWeb project brings together partners with different expertise.

[CRIDS](#) from University of Namur will provide expertise on the relevant legal issues.

[IDLab](#), [GhentCDH](#) and [MICT](#) of Ghent University will work on data enrichment, user engagement and evaluation and outreach to the research community, respectively.

KBR acts as project coordinator and will work on the development of the access platform and API and data enrichment.

	Top 50 FT MBA Ranking 2024 Cui Cui	2025-05-22
	Trump and Musk's discussion on X (12 August 2024) Valérie Schafer, Frédéric Clavert, Carmen Noguera	2025-03-22
	Royal Institute for Natural Sciences Belgium 2023-2024 Rolande Depoortere	2025-03-22
	Liechtensteinisches Volksblatt Social Media Accounts Philipp Wittwer	2025-03-22

Q&A

