



## Stream Learning for Multilingual Knowledge Transfer

<https://selma-project.eu/>

### D4.5 Demonstrator for use case one

Work Package	4
Responsible Partner	Priberam
Author(s)	Afonso Mendes
Contributors	Sebastião Miranda
Reviewer	Peggy van der Kreeft
Version	1.0
Contractual Date	31 March 2024
Delivery Date	28 March 2024
Dissemination Level	Public

## Version History

Version	Date	Description
0.1	14/03/2024	Initial version
0.2	18/03/2024	Internal review
0.3	19/03/2024	Final updates
1.0	25/03/2024	Publishable version

## Executive Summary

This document confirms that a demonstrator for use case one is available for user evaluation (available at <https://app.monitio.com>). SELMA use case one demonstrator incorporates into Monitio all the features and models developed within the project.

A detailed description of the work done for the UC1 demonstrator can be found in D1.4 (Final prototype Report). The evaluation of the demonstrator is reported in D5.3 (Final Evaluation Report).

# Table of Contents

- Executive Summary..... 3***
- 1. Demonstrator for UC1 – Monitio ..... 7***
- 2. Integration & Orchestration ..... 15***
  - 2.1 Monitoring .....15*
  - 2.2 RabbitMQ in Monitio.....16*
  - 2.3 Batch Maestro.....17*
- 3. Evaluation..... 19***
- 4. Conclusion..... 20***

# Table of Figures

**FIGURE 1** NAMED ENTITIES (LINKED TO WIKIPEDIA) IN THE MONITIO DOCUMENT PAGE, AS DETECTED BY THE NAMED ENTITY RECOGNITION AND LINKING MODEL DEVELOPED WITHIN SELMA ..... 7

**FIGURE 2** TRENDING ENTITIES ..... 8

**FIGURE 3** ENTITY NETWORK PAGE..... 9

**FIGURE 4** USER FEEDBACK COLLECTION UI FOR ENTITY LINKING ..... 10

**FIGURE 5** "STORYLINES" DASHBOARD FROM THE MONITIO PLATFORM, SHOWING THE CLUSTERS FROM THE CROSS-LINGUAL CLUSTERING MODEL DEVELOPED IN SELMA. ALTHOUGH AN ENGLISH TRANSLATION IS AVAILABLE FOR THE ARTICLES SHOWN, THEY ARE PRESENTED ABOVE IN THE ORIGINAL LANGUAGE TO EMPHASIZE THE MULTILINGUALITY OF THE PLATFORM..... 10

**FIGURE 6** IPTC TOPICS DETECTED ON A RUSSIAN DOCUMENT. THE TAGGING WAS DONE IN THE ORIGINAL LANGUAGE BY A MULTILINGUAL MODEL, WHERE TRANSLATION IS ONLY USED FOR SHOWING THE RESULT TO THE USER ..... 11

**FIGURE 7** MULTILINGUAL MULTI DOCUMENT SUMMARY OF A CLUSTER RELATED TO THE WAR IN GAZA. AFTER BEING SELECTED, THE SUMMARY SENTENCES ARE TRANSLATED TO ENGLISH ACCORDING TO THE PREFERENCES OF THE USER (COULD BE ANOTHER DESIRED LANGUAGE) ..... 12

**FIGURE 8** SPANISH VIDEO INGESTED BY MONITIO, INCLUDING AN AUTOMATICALLY EXTRACTED TEXT TRANSCRIPT, TOPIC DETECTION AND ENTITY RECOGNITION AND LINKING. A TRANSLATION TO ENGLISH IS ALSO AVAILABLE..... 13

**FIGURE 9** DIVERSITY FILTER COUNTS IN THE DWNEWS SCENARIO IN MONITIO FOR A PERIOD OF 30 DAYS BETWEEN OCT-22-2022 AND NOV-21-2022..... 14

**FIGURE 10** INTERNAL MONITORING BACK OFFICE SHOWCASING A FEW OF THE MONITORING JOBS THAT CHECK PERIODICALLY THE CORRECT FUNCTIONING OF THE PLATFORM..... 16

**FIGURE 11** MONITIO'S RABBITMQ INSTANCE OVERVIEW, SHOWING A FEW GLOBAL STATISTICS OF JOB QUEUE MESSAGE PROCESSING ..... 16

**FIGURE 12** MONITIO'S RABBITMQ INSTANCE, SHOWING A FEW OF THE JOB PROCESSING QUEUES, WHICH ARE ORGANIZED BY JOB TYPE (E.G., INDEXATION, VIDEO INGESTION, TRANSLATION) AND FEED GROUP SCENARIO (E.G, GLOBAL FOR ALL FEEDS, ES FOR SPANISH FEEDS)..... 17

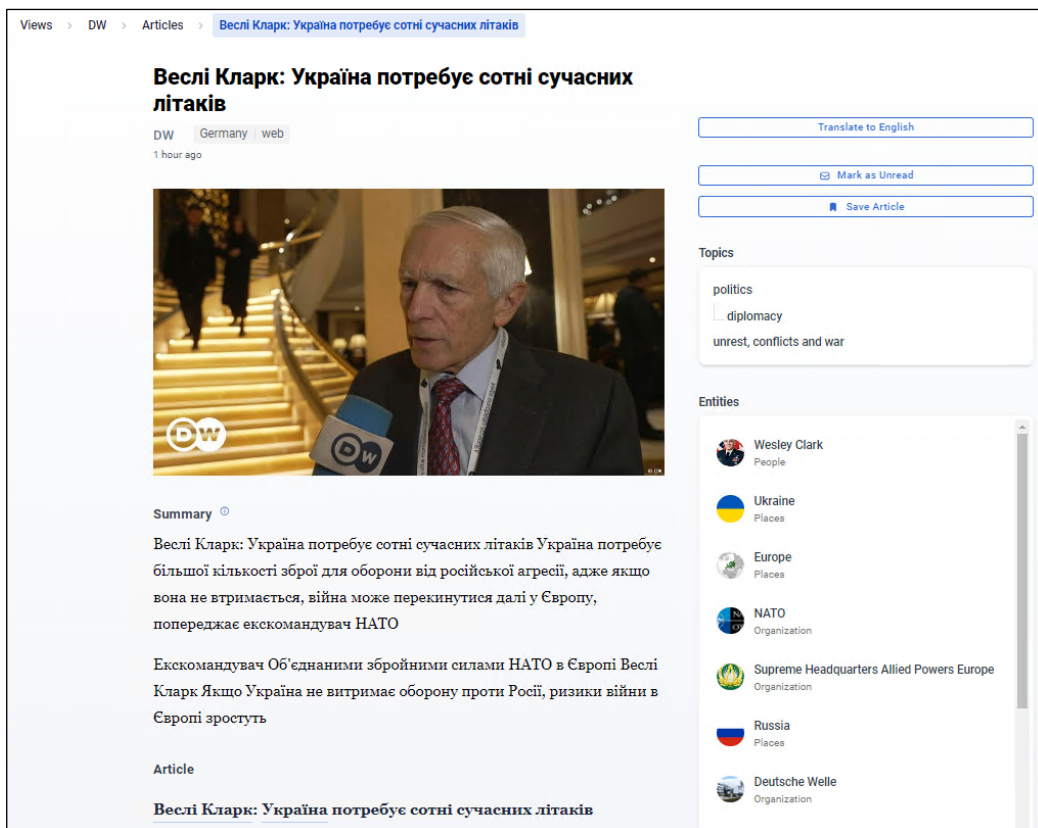
**Table of Tables**

*TABLE I MAESTRO BENCHMARKING RESULTS WHEN CHANGING THE PROCESSING BATCH SIZE AND IF THE DB CONTENTS ARE ENCRYPTED OR NOT. THIS IS JUST FOR ONE MAESTRO INSTANCE, MULTIPLE CAN BE ADDED TO SCALE THE SYSTEM TO MORE JOBS PER SECOND..... 18*

# 1. Demonstrator for UC1 – Monitio

The Monitio demonstrator for multilingual Media Monitoring - Use Case 1 - integrates into the platform the innovations done in the scope of the SELMA project. This report focusses on evidencing the features introduced into Monitio during the SELMA project, showing screenshots of the parts of Monitio UI where the features have been introduced. For additional information D1.4 has a deeper and more contextualized overview.

Figure 1 shows the integration into Monitio of the Multilingual Named Entity Recognition and Entity Linking against Wikipedia/Wikidata. Figure 1 shows a document with its detected Named Entities. Named Entities are used through the platform for search, filtering, trending views on entities, Entity network graphs, dashboards and reports.



The screenshot displays a news article interface. At the top, the breadcrumb navigation reads "Views > DW > Articles > [Веслі Кларк: Україна потребує сотні сучасних літаків](#)". The article title is "Веслі Кларк: Україна потребує сотні сучасних літаків", with subtext "DW Germany web" and "1 hour ago". A video thumbnail shows Wesley Clark speaking into a microphone. Below the video is a "Summary" section with text in Ukrainian. To the right, there are three buttons: "Translate to English", "Mark as Unread", and "Save Article". Below these are "Topics" (politics, diplomacy, unrest, conflicts and war) and "Entities" (Wesley Clark, Ukraine, Europe, NATO, Supreme Headquarters Allied Powers Europe, Russia, Deutsche Welle).

*Figure 1 Named Entities (linked to Wikipedia) in the Monitio Document page, as detected by the Named Entity Recognition and Linking model developed within SELMA*

Figure 2 shows the integration of the Named Entity Recognition and Linking model developed within SELMA, trending in a date range of 30 days.



**Figure 2** Trending Entities



Figure 3 shows Named Entities (linked to Wikipedia) and the connections found between them through document co-occurrence, in the Monitio Entity Network page. They are also shown on the right side of the filter pane, as detected by the Named Entity Recognition and Linking model developed within SELMA.

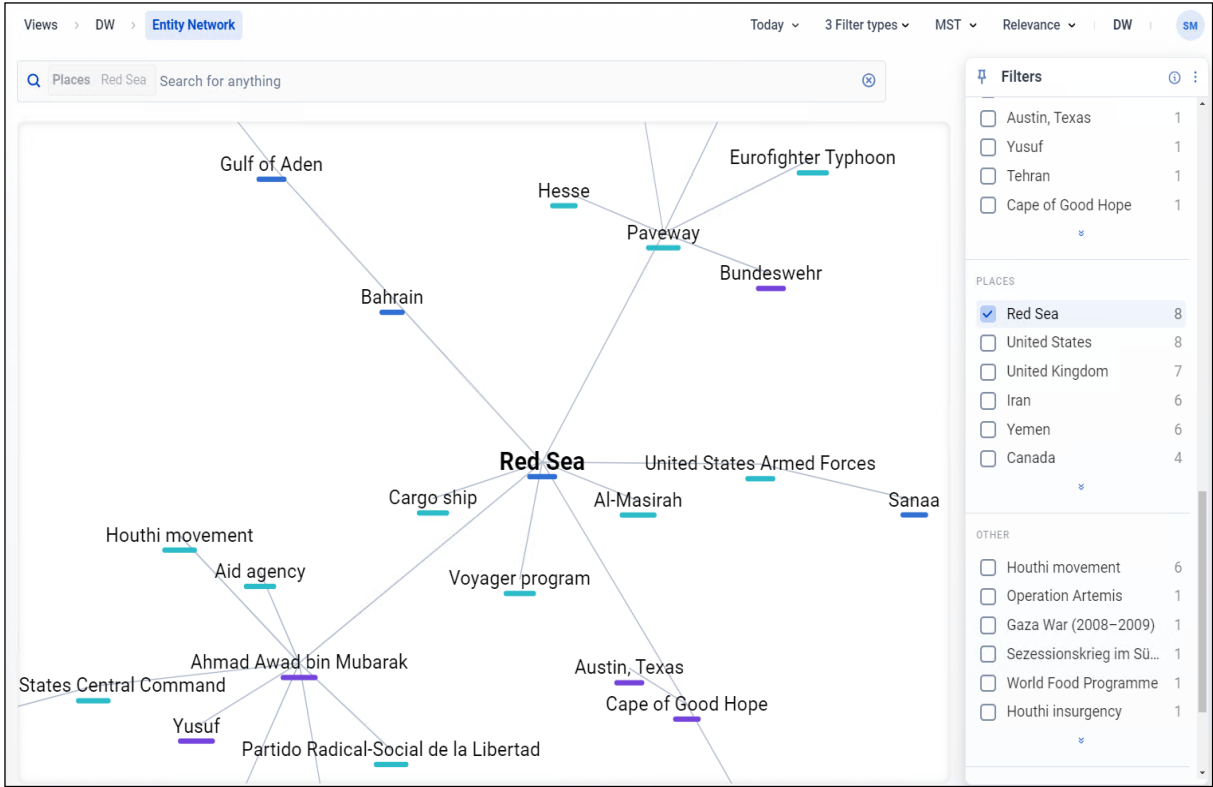


Figure 3 Entity Network page

Figure 4 shows the interaction for correcting Named Entity detection and provide feedback to the NER/EL models.

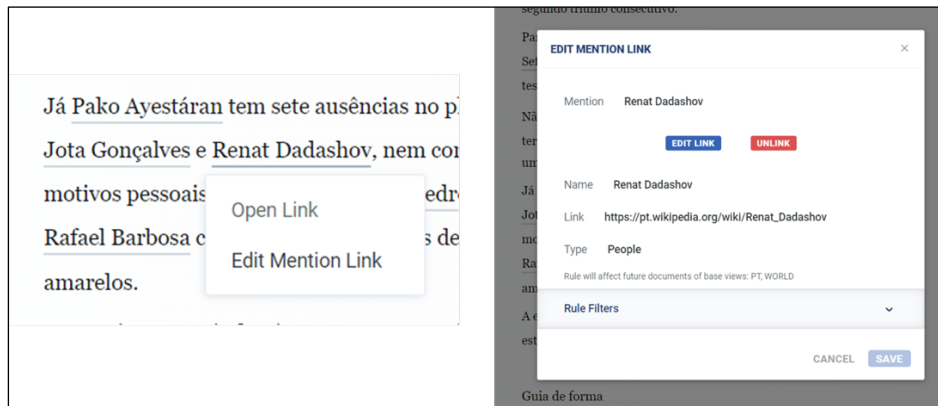


Figure 4 User feedback collection UI for entity linking

Figure 5 shows one of the integration pages on Monitio of the SELMA Multilingual News Clustering. Users can follow a story cluster over time, search over a cluster and see a generated summary based on the inner documents for a cluster.

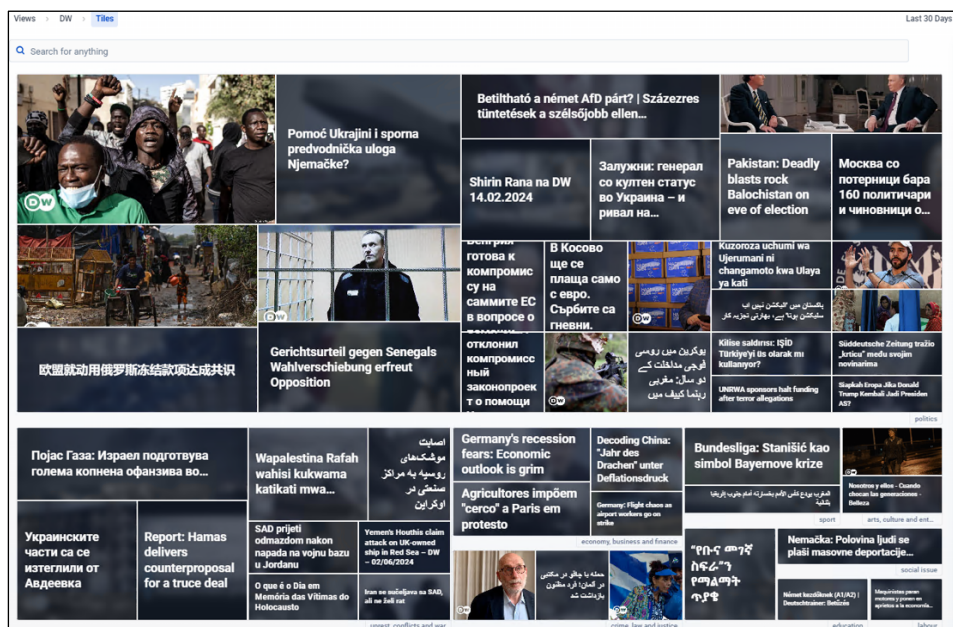


Figure 5 "Storylines" dashboard from the Monitio platform, showing the clusters from the cross-lingual clustering model developed in SELMA. Although an English translation is available for the articles shown, they are presented above in the original language to emphasize the multilinguality of the platform

Figure 6 shows the integration of the Multilingual Topic Detection using IPTC subject codes in Monitio. Topics are used through the platform for search, filtering, trending views on topics, Entity network graphs, dashboards and reports.




*Figure 6 IPTC Topics detected on a Russian document. The tagging was done in the original language by a multilingual model, where translation is only used for showing the result to the user*

Figure 7 shows the integration into Monitio of the Multilingual Multi Document Summarization. This summary presents selected sentences from the cluster documents that give an overview summary of the story covered by the cluster. The summary is presented in the user’s language of preference.


Views > GLOBAL > Tiles > Hundreds of truckloads of aid for Gaza stuck as more lives lost to malnutrition, some aid...

### Hundreds of truckloads of aid for Gaza stuck as more lives lost to malnutrition, some aid organizations say


70 articles from 43 feeds in 9 languages



**Hundreds of truckloads of aid for Gaza stuck as more lives lost to malnutrition, some aid organizations say**  
ABC News  
1 hour ago



**Más camiones de ayuda humanitaria entran en Gaza desde el inicio de la guerra - Informe**  
The Jerusalem Post  
Today at 11:24



**Unterstützung vor Ort: Hungersnot in Gaza: Können Spenden aus Deutschland helfen?**  
STERN  
Today at 10:21

**Summary**

- "Hundreds of truckloads of aid for Gaza stuck as more lives lost to malnutrition, some aid organizations say" ABC News
- "Israeli forces opened fire on people waiting for aid on Monday night in northern Gaza, eyewitnesses told CNN." Egypt Independent
- "Amidst the humanitarian crisis that causes deaths and suffering in the Gaza Strip, the United Nations World Food Programme (WFP) said on Tuesday that an aid convoy was blocked by the Israeli Army in Palestinian territory and then looted by several "desperate people." Folha de S. Paulo

*Figure 7 Multilingual Multi Document summary of a cluster related to the war in Gaza. After being selected, the summary sentences are translated to English according to the preferences of the user (could be another desired language)*

Figure 8 shows an ingested video on Monitio with its transcription and the rest of the NLP pipeline applied (NER/EL, Topic detection etc.).

**Terremoto político en Ucrania por tensiones entre la cúpula militar y política**

DW Germany | YouTube  
31/01/2024

Terremoto político en Ucrania por tensi...  
Watch Later Share

**Pulso en Zelenski y su jefe militar**

Watch on YouTube

**Video Transcription**

El presidente y su general el verano pasado, inspeccionando los sistemas antimisiles recién entregados. Una muestra pública de unidad. Pero a medida que la guerra se alargaba y una contraofensiva muy esperada contra los invasores rusos no conseguía torcer el rumbo, surgieron las tensiones. Al final de año, Zelensky dijo que el ejército le había presentado un plan, movilizar hasta 500.000 soldados más. Una medida políticamente arriesgada. Los soldados en primera línea llevan meses, algunos de ellos años, de servicios sin descanso. La opinión

**Translate to English**

**Mark as Unread**

**Save Article**

**Topics**

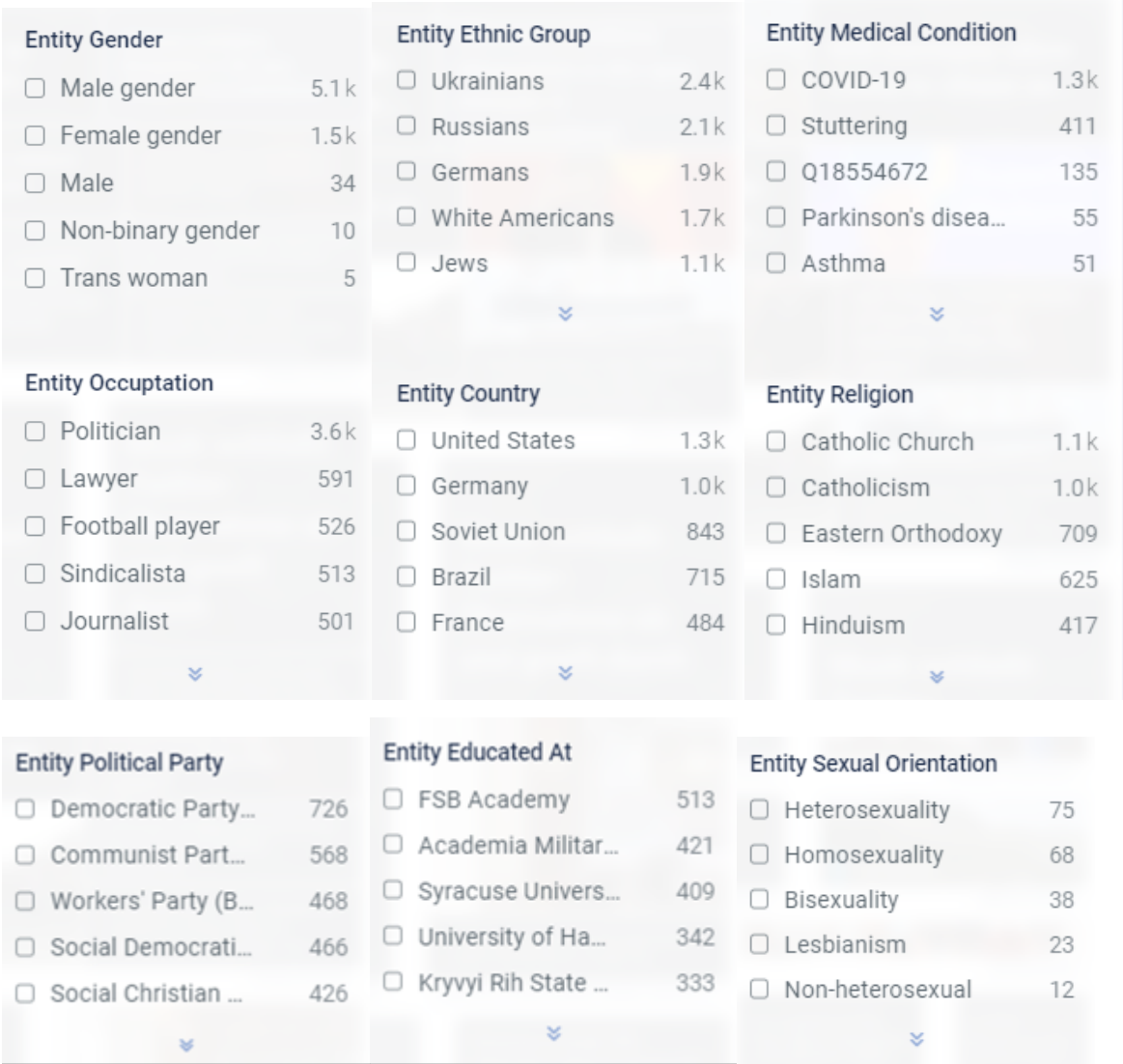
- unrest, conflicts and war
- politics
  - defence
  - armed Forces

**Entities**

- Volodymyr Zelenskyy (People)
- Ukraine (Places)
- Verkhovna Rada (Organization)
- The Kyiv Independent (Organization)
- Kyiv (Places)

*Figure 8 Spanish video ingested by Monitio, including an automatically extracted text transcript, topic detection and entity recognition and linking. A translation to English is also available*

Figure 9 shows the diversity filters integrated into Monitio, these are also available for external integration at DW to collect statistics on their production.



**Figure 9** Diversity filter counts in the DWNEWS scenario in Monitio for a period of 30 days between Oct-22-2022 and Nov-21-2022



## 2. Integration & Orchestration

Monitio takes advantage of the worker management and scalability of DockerSpaces and the SELMA Maestro orchestrator (See WP4 deliverables).

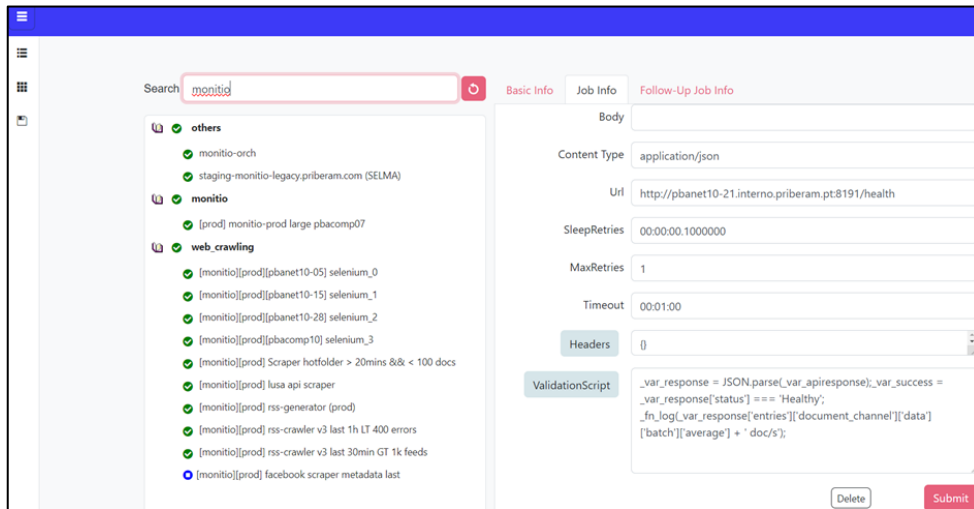
The models/components integrated in Monitio are:

- SELMA Named Entity Recognition and Entity Linking
- SELMA Entity User Correction
- SELMA News Clustering
- SELMA Topic Detection
- SELMA Summarization
- SELMA Orchestration (Maestro, Docker Spaces)
- Open Source Machine Translation (Meta's m2m\_100)
- Open Source Speech to Text (Whisper)

All these components have been packaged in docker containers and integrated as RabbitMQ workers, which can be launched with Docker Spaces.

### 2.1 Monitoring

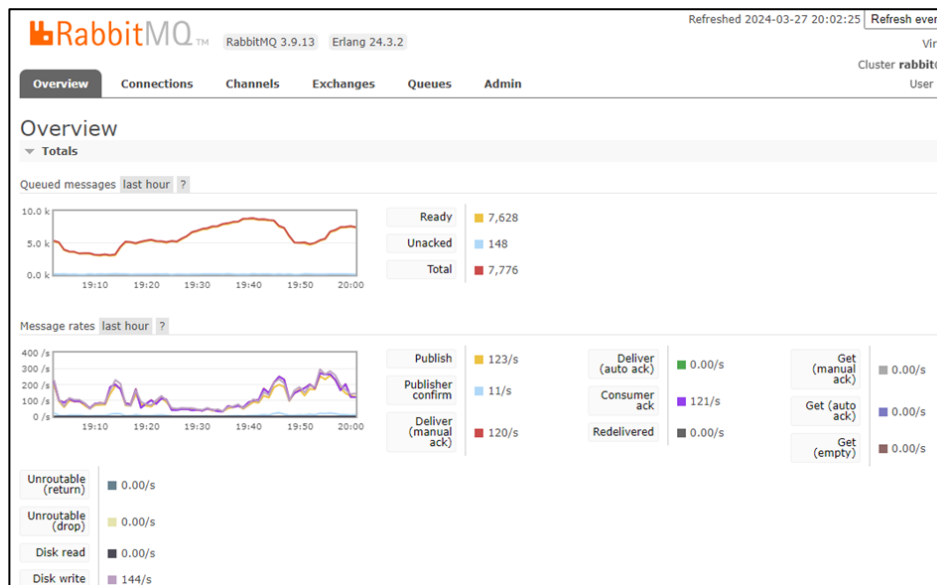
The Monitio backend, scraping and orchestration are actively monitored using a tool which leverages the same javascript job execution technology as Maestro (See WP4 deliverables), but for monitoring jobs instead of NLP processing jobs. The same system is used in plain X.



*Figure 10 Internal monitoring back office showcasing a few of the monitoring jobs that check periodically the correct functioning of the platform*

## 2.2 RabbitMQ in Monitio

RabbitMQ is used to manage pending processing jobs in the Monitio orchestration pipeline. Following are two images which show a snapshot of a few of the processing queues, which are organized by job type and a feed group.



*Figure 11 Monitio's RabbitMQ instance overview, showing a few global statistics of job queue message processing*



RabbitMQ 3.9.13 Erlang 24.3.2 Refreshed 2024-03-27 20:03:16

Overview Connections Channels Exchanges **Queues** Admin

Queues  
All queues (101, filtered down to 88)

Page 1 of 1 - Filter: Priberam  Regex ?

Overview				Messages			Message rates		
Name	Type	Features	State	Ready	Unacked	Total	incoming	deliver / get	ack
Priberam.Indexation_GLOBAL	classic	D Args	running	6,615	1	6,616	18/s	2.2/s	2.2/s
Priberam.Indexation_ES	classic	D	running	539	1	540	19/s	2.2/s	2.2/s
Priberam.Translation_GLOBAL	classic	D	running	234	6	240	0.80/s	0.20/s	0.00/s
Priberam.Indexation_ES_lite	classic	D	running	37	1	38	19/s	12/s	12/s
Priberam.Indexation_GLOBAL_lite	classic	D	running	12	1	13	18/s	11/s	11/s
Priberam.NearDuplicates_ES	classic	D	running	7	1	8	0.40/s	2.6/s	2.8/s
Priberam.NearDuplicates_GLOBAL	classic	D	running	3	1	4	1.6/s	3.2/s	3.4/s
Priberam.IptcTopics	classic	D	running	2	1	3	2.6/s	3.0/s	3.2/s
Priberam.icd9	classic	D	idle	0	0	0	0.00/s	0.00/s	0.00/s
Priberam.hnner.PHARMA	classic	D	idle	0	0	0	0.00/s	0.00/s	0.00/s
Priberam.VideoIngestion	classic	D	idle	0	0	0	0.00/s	0.00/s	0.00/s
Priberam.Translation.UK	classic	D	idle	0	0	0	0.00/s	0.00/s	0.00/s
Priberam.Translation.TECH	classic	D	idle	0	0	0	0.00/s	0.00/s	0.00/s
Priberam.Translation.MEDICINE	classic	D	idle	0	0	0			

*Figure 12 Monitio's RabbitMQ instance, showing a few of the job processing queues, which are organized by job type (e.g., Indexation, Video Ingestion, Translation) and feed group scenario (e.g, GLOBAL for all feeds, ES for spanish feeds)*

### 2.3 Batch Maestro

Additional stress tests have been performed on Maestro to increase processing throughput. We have concluded that changing the maestro component to work in a batch mode greatly improved performance, as shown in the figure below. Note that it is possible to scale to multiple Maestro components to cope with increasing Job/s requirements, but it was important to improve the performance of a single maestro instance to optimize resources. We have also benchmarked the cost of adding encryption to the database, which is especially relevant when processing sensitive data in Monitio or for the case of plain X where we deal with user uploaded content.

We have also extended the integration mechanisms of Maestro with job workers to improve the system's extensibility and scalability.

Encryption		Consumer Type	Batch Size	Concurrency	Total Time (ms)	Jobs/s	Avg. Time (ms)	Processing Time %
total jobs processed: 4900								
	SINGLE	1	n/a	111323	44.016	22.719	103%	
	SINGLE	50	n/a	146389	33.472	29.875	135%	
	SINGLE	50	n/a	108047	45.351	22.050	100%	
	BATCH	1	1	73180	66.958	14.935	68%	
	BATCH	1	2	71996	68.059	14.693	67%	
	BATCH	5	1	41531	117.984	8.476	38%	
	BATCH	50	8	32394	151.263	6.611	30%	
	BATCH	100	8	32057	152.853	6.542	30%	
	BATCH	200	1	31986	153.192	6.528	30%	
	BATCH	50	4	68977	71.038	14.077	64%	
	BATCH	50	4	31977	153.235	6.526	30%	
	BATCH	100	4	31305	156.525	6.389	29%	
	BATCH	50	1	31288	156.610	6.385	29%	
	BATCH	100	1	30832	158.926	6.292	29%	
	BATCH	100	1	71031	68.984	14.496	66%	
	BATCH	100	4	68741	71.282	14.029	64%	

*Table 1 Maestro benchmarking results when changing the processing batch size and if the DB contents are encrypted or not. This is just for one Maestro instance, multiple can be added to scale the system to more jobs per second*

### 3. Evaluation

The demonstrator evaluation was carried out within WP5 and reported in deliverable D5.3 (Final Evaluation Report). The demonstrator is being used at DW and the Monitio API is being used by the Podcast Creator as well as the Diversity Indicator (both at DW). The demonstrator has also been shown or given access to potential clients and members of the user group. Priberam has already managed to engage the first clients. During the project, user feedback has been continuously integrated in the development and research cycles, either to improve the underlying ML models or the UI itself.

At time of writing, the demonstrator is able to ingest 300K documents per day and apply the full NLP pipeline.

## 4. Conclusion

The demonstrator has integrated the multilingual and user feedback components developed within the SELMA project. The use of the demonstrator by the testers and the user group provided excellent feedback on usability and model performance as perceived by the final user. Monitio reached a new level in its Machine Learning models with an emphasis on its multilinguality.