DATAWEEK

JOIN.LEARN.SHARE.GET VALUE

SELMA Shaping (Big Data) speech and text technologies for media monitoring and the newsroom

EUHUBS4DATA

BDV BIG DATA VALUE ASSOCIATION

SELMA

SELMA

Shaping (Big Data) speech and text technologies for media monitoring and the newsroom (01.2021-12.2023)

BDVA: "Data meets infrastructure at the Edge", 14.06.2023

Kay Macquarrie

priberam

LATVIJAS UNIVERSITĀTE

Fraunhofer

AVIGNON UNIVERSITÉ

Deutsche Welle

H2020 // Call: H2020-ICT-51-2020-RIA // Number 957017

Two Use Cases

Find content

Advanced Content Analysis including Broadcast Monitoring & Diversity Data; Press Agency Analysis

Media Monitoring

Produce content

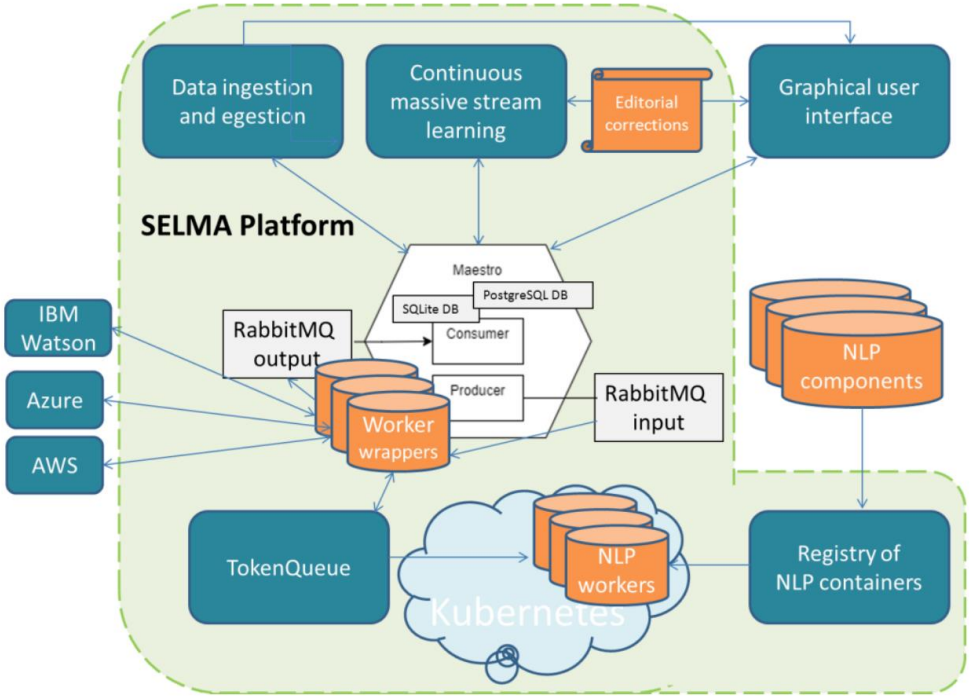
News Podcast Creation, the Video Subtitling and the Video Voice-over

Media Production

H2020 // Call: H2020-ICT-51-2020-RIA // Number 957017

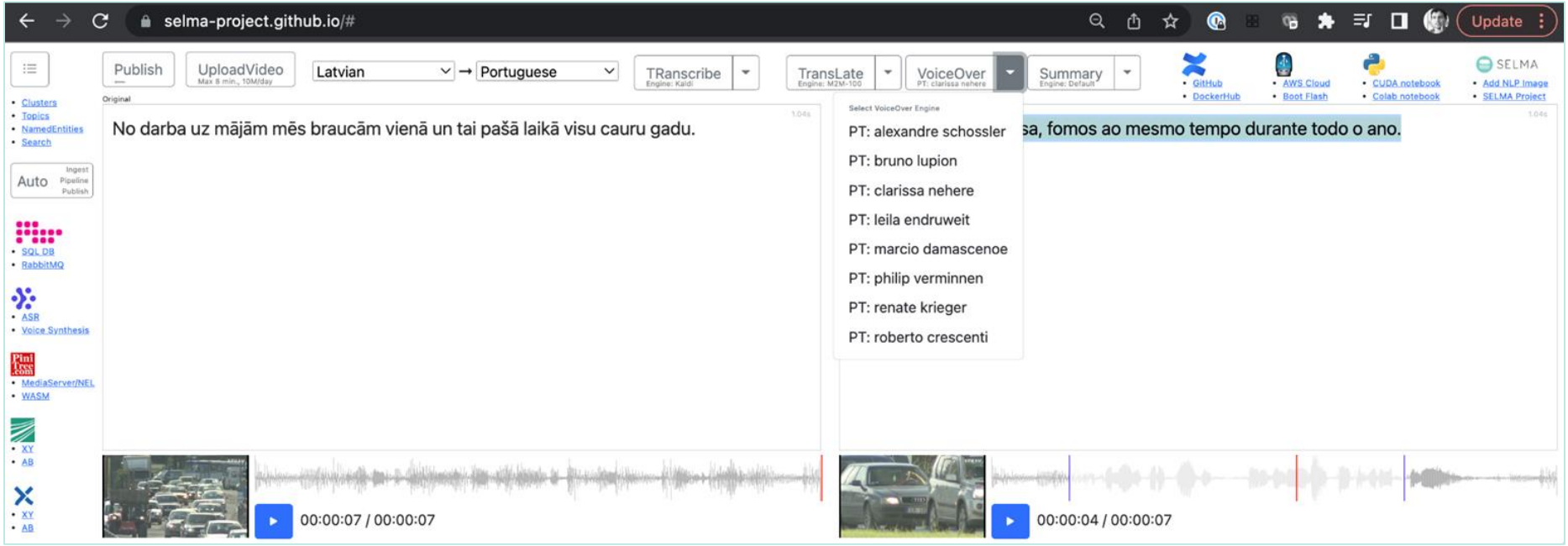
SELMA

SELMA output: scalability



SELMA

SELMA Open-Source Platform (beta)

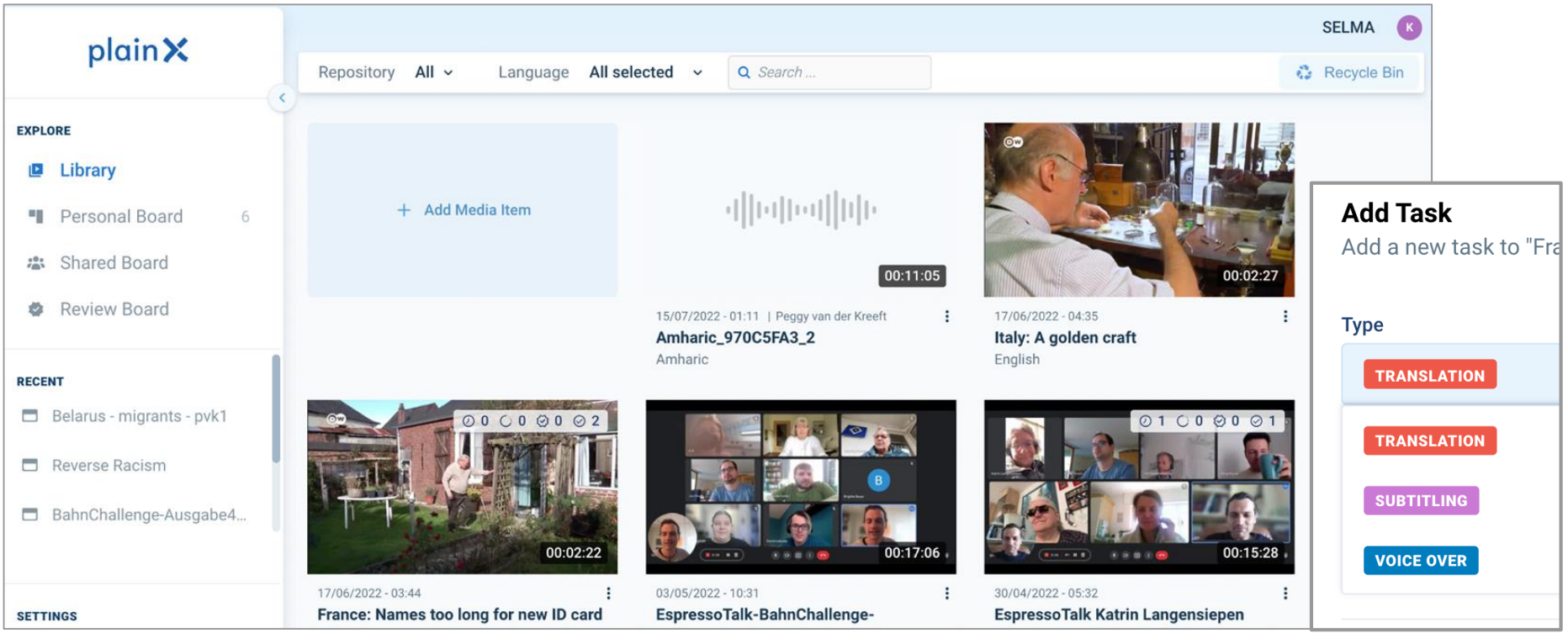


H2020 // Call: H2020-ICT-51-2020-RIA // Number 957017

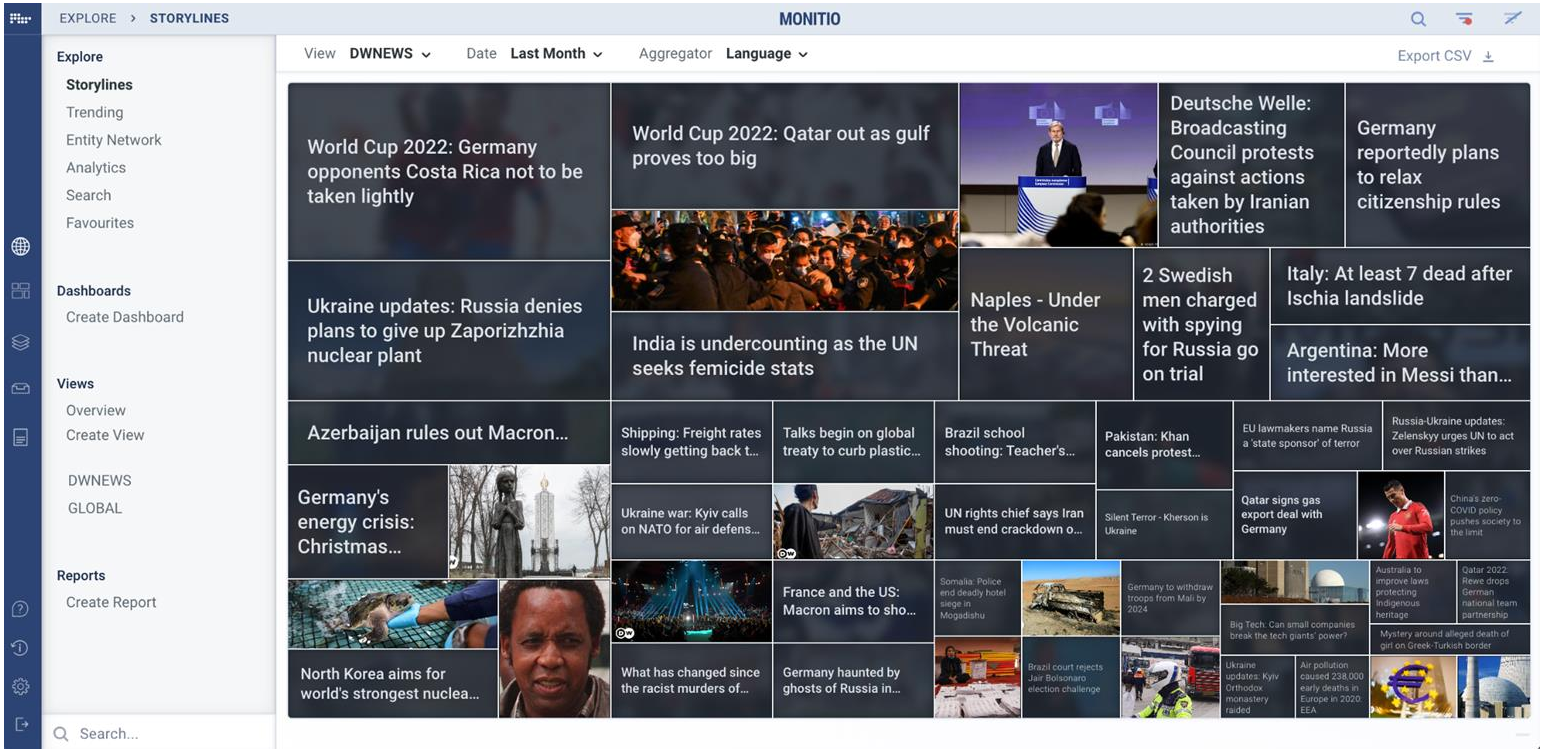
SELMA

plain X (Use Case)

NLP tool geared for multilingual content production



Monitio (Use Case)



H2020 // Call: H2020-ICT-51-2020-RIA // Number 957017

SELMA

Contributions to Innovation Radar

#1	**plain X**	A novel tool for easy transcription, translation, subtitling & voice-over
#2	**Scalability Module**	Highly scalable SELMA NLP orchestration platform for processing extreme volumes
#3	**Podcast Creation**	Create News Podcasts using NLP analysis and summarization

Contribution to Data Spaces

- Build an open-source platform for Big Data Analysis (Orchestration = integrates NLP processes)
- Multilingual Adaption Workflow for ASR, MT and Voice Synthesis (open-source modules)
- Proprietary advanced multilingual editorial adaptation workflow (plain X)
- Proprietary multilingual media monitoring platform (Monitio)
- DW provides training and test data from all its 32 languages

Collaboration on technical aspects

- NER / NEL modules for a large number of languages
- DW Benchmarking of NLP processes (ASR, MT, speech synthesis) for many languages
- Experimentation with speech to text translation and speech to speech translation

Future Work

- How Language Space & Data Space will be intertwined
- Transfer Learning from HRL to LRL (High resourced languages to Low resourced languages) or between domains
- Data Resources Gap: Get a better grip of what data is available or get native speakers involved (costly)
- Make efficient use of multilingually processed data (broadcast, other data)

H2020 // Call: H2020-ICT-51-2020-RIA // Number 957017

SELMA

THE DATA IS OUT THERE

DATAWEEK

JOIN.LEARN.SHARE.GET VALUE

Thank you!

Basics

RIA, EU Horizon 2020

Runtime: January 2021-December 2023

Grant Agreement 957017

selma-project.eu

twitter.com/selma_project

Coordination: Kay Macquarrie, DW

Contact: info@selma-project.eu

EUHUBS4DATA

BDV BIG DATA VALUE ASSOCIATION

SELMA