Research and Innovation Action (RIA) H2020 – 957017

SELMA

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D7.3 Final Periodic Report

Work Package	7
Responsible Partner	DW
Author(s)	Kay Macquarrie (DW), Tim Koch (DW)
Contributors	Guntis Barzdins (IMCS), Afonso Mendes (Priberam), Yannick Estève (LIA), Peggy van der Kreeft (DW)
Version	1.0
Contractual Date	31 March 2024
Delivery Date	28 March 2024
Dissemination Level	Public

Version History

Version	Date	Description
0.1	07.03.2024	First draft
0.2	12.03.2024	Updated draft
0.3	14.03.2024	Integrated contributions from all partners
0.9	16.03.2024	Draft version
1.0	27.03.2024	Final (draft) version for submission

Table of Contents

Explanation of the work carried out by the beneficiaries and overview of progress.

Very large amounts of multilingual information in the form of data are all around us and are growing rapidly. With the emergence of LLMs / Large Language Models (open AI launched Chat GPT at the end of 2022), the AI and NLP world changed significantly. In SELMA we made use of the developments around LLMs, for instance to generate tag descriptions for selected keyword pairs. Still, the potential to fully take advantage of digital content streams based on machine learning had remained widely untapped even with the existence of LLMs.

SELMA tackled these potentials from two sides in the past three years: by significantly advancing multilingual language technologies from a research perspective and by integrating concrete technological improvements into components, platforms and prototypes which to some extent will be available open source for the public and (the media) industry.

A focus of the SELMA work was set into a unified approach to multilingual media monitoring and content production by leveraging and contributing to advances in deep learning, in particular in multilingual language modeling, knowledge transfer and language transfer. With a consortium of three research institutes/universities (Fraunhofer Gesellschaft IAIS, University of Avignon and University of Latvia/IMCS) and two industry/broadcasting companies (Priberam, an SME, and Deutsche Welle, a large international broadcaster) significant progress into “shaping speech and text technologies for media monitoring & the newsroom” had already been made in the first period of the project (01.01.2021-30.06.2022).

In the second period of the project (01.07.2022-31.03.2024 including a 3-month-extension), significant research results in the field of language technology could be made and integrated into the platforms (UC0, UC1 and UC2) and Use Case Applications / Prototypes (Podcast Creator, Diversity Application, M-PHANTOM, Diarization, DW Speaker, DW Summarizer). All objectives and KPIs were met and a great part of the developed software and components as well as the SELMA open-source platform could be released as public domain. The plain X platform evolved into a product, was rolled out at Deutsche Welle and could gain first clients.

The Monitio platform was enriched with a new NLP orchestration pipeline and new multilingual NLP analyses based on state-of-the-art AI methods, thus making it more scalable and avoiding a language bottleneck of translating the content into English.

SELMA was organized around 3 Use Cases and 5 Use Case Applications

> **Use Case 1: Media Monitoring** - based on the Monitio Platform. This use case analyzes and filters (very) large amounts of media data streams and comprises two use case applications: Advanced Content Analysis, including Broadcasting, and Diversity.
>
> **Use Case 2: News Production** - based on the plain X Platform. This use case provides an editorial production workflow for NLP processing tasks such as transcription, translation and voice-over. It has resulted in the development of three use case applications: News Podcast Creation, Video Subtitling and Video Voice-Over.
>
> **Use Case 0: SELMA OSS** - was introduced and developed for testing SELMA NLP components and models in an integrated open-source platform.

Objectives

The overall aim was to "build a continuous (...) deep learning platform using extreme analytics, transfer learning and advanced natural language processing technologies". We have made significant progress in many of the eight objectives during the first and the second period of the project.

Objective 1: Massive processing of audio/video/text data streams

A core objective of SELMA was to process extremely large amounts of audiovisual and textual data. In the first period, the infrastructure was developed, and relevant news data streams were identified and fed into the system. Environments were established for the two main use cases. The open-source platform, capable of processing up to 10 million items per day, was constructed. See deliverable D4.2 Initial Platform Release with the primary NLP pipeline for details.

In the second period, the open-source platform (including the scaling component) was integrated into the platforms. The final SELMA open-source platform was made available

online as well as a downloadable version that can be locally configured and personalized. Both versions are publicly available on GitHub.

Related milestones

P1- MS2 Release of Architecture Agreement (M10)
MS4 Release of Platform Release v1(M15)
P2 - MS10 Scalability Testing (M34)

Objective 2: Unsupervised multilingual language models in a shared space for 30 languages

This objective concerned the research and development of new methods for training deep learning unsupervised languages models in 30 (+) languages. It mainly concerns work done in WP2 and WP3.

In P1, 17 academic papers were published and significant methodical improvements for various downstream tasks (including entity recognition and linking, topic labelling, clustering, summarization, transcription and translation) could be made and integrated into the platforms.

In P2, 14 academic papers were published and, again, significant improvements for the downstream tasks could be made (cf. KPI in D6.6).

Related milestones

P1 - MS3 Release of Initial Prototypes (M12)
P2 - M12 Final Prototypes (M39)

Objective 3: Knowledge transfer across tasks and languages

This objective aimed at researching new methods for knowledge transfer across tasks and languages with asymmetrical amounts of resources available among different languages and tasks.

In P1, initial progress of transfer learning has been made and described in the technical deliverables (D2.2, D2.3 and D3.3).

In P2, very good results for knowledge transfer across languages development could be reached (D2.7, D2.8, D3.7, D3.8). Various methodologies were researched and applied to different tasks like Named Entity Recognition, topic detection, ASR and speech modeling, etc. In the end, many of the models were integrated into the platforms with high TRL levels.

Objective 4: Enable media monitoring analytics for decision-making

This objective aimed to improve decision-making processes by developing novel data analytics and visualization methods. The target group are media monitoring analysts but also any global end-user.

In P1, results were integrated into Monitio from the following WP2/3 components: Multilingual News Clustering, Multilingual Topic Detection, and Rule-Based Entity Correction. More recently, the following SELMA research results related to the Multilingual Entity Linking component were added to Monitio. Regarding the Advanced Content Analysis application goals, the project has set up a Wikidata processing pipeline to retrieve relevant properties of known entities (D1.2).

In P2, further results were integrated into the Monitio platform which led to even better results. The Monitio API is used by the Podcast Creator (cf. Objective 5) and for the Diversity Balance Indicator. The better language coverage across models and the increased processing capacity provided by the orchestration pipeline improved significantly this kind of analysis and its coverage available to the end-user.

Related milestones

P1 - MS1 Release of Requirements Analysis (M6)
MS5 User Evaluation 1 (M15)
P2 - MS15 Final Platform Release M39
MS9 User Evaluation 2 M24
MS13 User Evaluation 3 M39
MS15 Final Platform Release M39

Objective 5: Enable multilingual content production workflow

This objective was to provide a content production workflow by leveraging multilingual transcription and translation models trained within SELMA. The target group is journalists and editorial production teams.

In P1, the plain X user interface has been developed based on SELMA's UC2 requirements (see D1.2) and results from WP2/3, including Automatic Speech Recognition, Speech Translation and Punctuation modules were integrated into plain X; in the backend, plain X is using the SELMA orchestration to schedule and execute NLP jobs (see D4.1 and D4.2). The SELMA OSS platform (UC0) was developed and provides a basic platform to do transcription, translation, and voice-over tasks in selected languages. Work had started on the Podcast Producer tool, a separate application whose purpose is to use UC0's speech synthesis modules to support the semi-automated creation of news podcasts.

In P2, plain X has evolved into a product with a roll-out into Deutsche Welle with an ever-increasing number of content production (transcription, subtitling and translation), the launch of a product website (plainx.com) and first external clients. Results of WP2 and WP3, including speech synthesis modules, have been integrated. The SELMA open-source platform was developed and made publicly accessible in its final version (both web-based as well as a downloadable local version). The SELMA OSS API is used by the Podcast Creator, DW Summarizer and DW Speaker.

Related milestones

P1 - MS1 Release of Requirements Analysis (M6)
MS5 User Evaluation 1 (M15)
P2 - MS15 Final Platform Release M39
MS9 User Evaluation 2 M24
MS13 User Evaluation 3 M39
MS15 Final Platform Release M39

Objective 6: Fine-tune deep learning models from user feedback

This objective aimed to improve and fine-tune deep learning models from user feedback based on novel deep learning methods.

In P1, an end-to-end model for named entity recognition from speech without paired training data was built in the framework of the SELMA project. This enabled the project to investigate novel ways to generate massive amounts of training data for the post-editing task (D3.3). Initial release of post-editing and user feedback capabilities for details was done.

In P2, SELMA has integrated post-editing and user feedback capabilities into the platforms. M-PHANTOM was developed for ingesting annotations from users gained in an editorial news production setting into the learning models to continuously improve the use of the SELMA platform.

Objective 7: Sustainable exploitation of the SELMA platform

This objective relates to the incorporation of SELMA results into media environments such as Deutsche Welle.

In P1: With the set-up of the SELMA open-source platform many technological developments within SELMA were made available for public use. Outcomes of WP2, WP4 and WP4 have been integrated into versions of both the plain X as the Monitio platform.

In P2, major exploitational development have been made, including the clearance of IPR for 26 components (half of them being released as open source), the set-up of plain X the product framework, and several workshop events with the aim to promote and exploit the platforms and prototypes in the media domain. Negotiations with initial external clients for plain X were started.

Related milestones

P2 – MS7 Components Release v1 M24
MS14 Final Components Release M39
MS16 Sustainability Plan M39

Objective 8: Dissemination and communication of the SELMA project outputs

This objective mainly concerned work done in WP6.

In P1, SELMA set up a user group and engaged with all relevant stakeholders in the language technology and innovation chain, including broadcasters, commercial players, EU agencies and

the relevant research communities. Measurable progress was made through various project presentations and engagement in external events as well as several project publications (D6.2).

In P2, SELMA was present at various events (workshops and conferences including two user days, one in Bonn and the other one in Avignon, and two user group events) and made many publications including an award / prize winning one from Cambridge University (for the paper ICASSP 2023 « Federated Learning for ASR based on Wav2vec 2.0 » from LIA). The website (plus social media channels) was updated regularly with articles and posts around NLP and Big Data topics (D6.6).

Related milestones

P2 - MS6 User Day M1
MS11 User Day 2 M35

Explanation of the work carried out per WP

WP1 Requirements and Prototyping

Work package number	WP1	**Lead beneficiary:**	Priberam
Work package title	Requirements and Prototyping		
Start month – End Month	M1 – M39		

WP leader: Priberam

Participating Partners: DW, PRIB, IMCS, LIA, FhG

Task T1.1 Use Case Description and Requirements

SELMA's NLP research on transfer learning, user feedback learning and stream learning has been applied into three main use cases, Multilingual Media Monitoring (UC1), Multilingual News Content Production (UC2) and SELMA NLP Service Orchestration (UC0), through the

development of different software prototypes. Use-cases UC1 and UC2 and the corresponding requirements are discussed in detail in deliverable D1.1, whereas Use-case UC0 is discussed in deliverable D1.2. Deliverable D1.1 also specifies use-case scenarios for UC1 Advanced Content Analysis and Press Agency Analysis, and News Podcast Creation, Video Subtitling and Video Voice-Over for UC2. The progress made in implementing the Requirements was reported periodically in deliverables D1.2, D1.3 and D1.4.

Task T1.2 Wireframing

This task supported tasks T1.3 and T1.4 by developing wireframes and mockups to drive the development of the UC1 (https://app.monitio.com) and UC2 (https://app.plain-x.com) prototypes. Deliverables D4.5 and D4.6 document the release of these two demonstrators, whereas a more complete description of the main Use-Case prototypes and other component prototypes is reported on D1.4.

Task T1.3 Multilingual Media Monitoring Prototype

Within Use Case 1 (UC1), we have integrated the results from the SELMA research tasks into the *Monitio* product, a Media Monitoring platform under development by Priberam, available at https://app.monitio.com. Some of these research results are improved models using transfer and stream learning (see D2.7 for technical details), while others are changes to the platform to allow learning from user feedback (see D3.7 for technical details). Another prototyping effort integrates the SELMA NLP Service Orchestration (UC0) as *Monitio*'s job orchestrator, which allows *Monitio* to scale (see D4.1-D4.4). The SELMA models integrated in Monitio are Named Entity Recognition and Entity Linking, Entity User Correction, News Clustering, Topic Detection and Summarization (see WP2 deliverables). Other prototyping efforts and improvements done within SELMA are the integration of OSS translation (META's m2m100) and transcription (OpenAI's Whisper) models, inclusion of A/V processing in the pipeline, Entity Diversity Data ingestion for the Diversity Use-Case, and integration of grounded Large Language Models (LLMs) for Explainability of analytics results. Implementation details in D1.4.

Task T1.4 Multilingual News Production Prototype

Within Use Case 2 (UC2), we integrated results from SELMA into the *plain X* product, a Multilingual News Media Content Production platform under development by Priberam and Deutsche Welle, available at https://app.plain-x.com. plain X has been mostly developed within SELMA, started from a first prototype named "news.bridge".

To achieve SELMA's research goals on learning from user feedback, *plain X* stores original transcripts and user-edited versions such that this data can be used to improve models (if approved by the corresponding user).

plain X is using the SELMA orchestration to schedule and execute NLP jobs (D4.4), the ASR, Speech Translation and Text to Speech SELMA models (D3.7), and the MT model integrations (D4.4). In the case of *plain X*, the SELMA orchestration allows to execute NLP jobs not only from self-hosted APIs (e.g., the ones developed within SELMA through DockerSpaces – see D4.4) but also from many cloud providers (Azure, Google, etc).

A "Podcast Creator" use-case was also developed based on a workflow observed in DW's Brazilian language department. The use case's goal is to increase the workflow's efficiency by supporting the journalist in the production of daily audio news bulletins through SELMA, using UC0, UC1 and UC2's APIs. The Podcast Creator Prototype was also later broken down into separate standalone apps "DW Summarizer" and "DW Speaker" for executing the sub tasks of summarization and text to speech. Details in D1.4.

WP2 Continuous Massive Stream Learning

Work package number	WP2	**Lead beneficiary:**	Fraunhofer
Work package title	Continuous Massive Stream Learning		
Start month – End Month	M1 – M39		

WP leader: Fraunhofer

Participating Partners: DW, PRIB, IMCS, LIA, FhG

Task T2.1 Cross-lingual Stream Representations

We significantly enhanced our foundational cross-lingual stream representation models in the second phase. Our refined approach now better handles the intricacies of cross-lingual knowledge transfer and entity embedding creation, using advanced pipelines that iteratively refine contextual word embeddings. This process, starting with mean pooling and potentially incorporating Wiki data, now utilizes Wikipedia contexts for generating more precise multilingual embeddings, optimizing them for tasks like entity linking.

We've expanded our model's linguistic capabilities to include up to 40 languages, demonstrating improved resilience and consistent performance across this broader spectrum. This expansion, alongside the integration of cutting-edge methods and technologies, has markedly advanced our cross-lingual application development, enhancing knowledge transfer across languages.

Additionally, we've made significant strides in memory efficiency, adopting the bfloat16 format for training and storing embeddings. This development enables the management of approximately 20 million Wikipedia entities, setting a new standard for scalability and efficiency in the field.

Task T2.2 Named Entity Recognition and Linking

We have significantly progressed in enhancing named entity detection and linking within news streams. Our focus on refining models, especially Hierarchical Nested Named Entity Recognition (HNNER) and example-based NER, has led to notable improvements in precision and efficiency. A key advancement is a novel architecture based on HNNER, featuring enhanced computational efficiency and an attention-only mechanism, enabling efficient cross-ontology training and application.

Leveraging contextual embeddings has significantly improved our entity recognition processes while exploring k-nearest neighbors (kNN) methodologies has increased model adaptability to evolving datasets without frequent retraining. A major achievement has been developing a unified model capable of optimal performance across languages and ontologies, illustrating our commitment to scalable and efficient multilingual NER solutions.

Our training strategies, including replacing traditional architectures with attention mechanisms, promise further enhancements in model performance across diverse datasets. These efforts mark

a significant step towards overcoming the challenges of multilingual and multi-ontology entity recognition and linking, ensuring scalable, efficient, and precise solutions for SELMA applications.

Task T2.3 Story Segmentation

We have made notable progress in speaker analysis, particularly in speaker clustering and diarization, marking a significant advance from our initial phase. Our work has expanded to include sophisticated speaker identification, accurately matching voice samples to labeled speaker models, and enhancing speaker diarization tasks crucial for segmenting speaker information in lengthy news content. To achieve this, we've developed a robust system that accurately identifies speaker change points, grouping speech segments by speaker traits.

Our investigation into text-independent speaker recognition has focused on identifying speakers by speech characteristics, employing speaker embeddings from unconstrained utterances. We've explored advanced x-vector-based architectures and time-delayed neural networks (TDNNs), significantly enhancing our speaker-related tasks' capabilities. Incorporating the ECAPA-TDNN architecture has allowed us to overcome previous limitations, extending temporal attention to focus more granularly on speaker-specific features and improving our ability to segment speakers accurately, regardless of linguistic content.

Additionally, integrating ResNet modules for multi-scale feature processing and employing multi-layer feature aggregation (MFA) have optimized performance and reduced model complexity. These enhancements and optimizing the AAM-soft-max loss for training have refined our approach to calculating cosine distances between speaker embeddings, offering a more sophisticated method than traditional PLDA.

Task T2.4 Online News Classification and Clustering

Our second phase progress has significantly refined our approaches to classifying and clustering online news content, leveraging the International Press Telecommunications Council (IPTC) taxonomy for a nuanced categorization across a wide linguistic spectrum. Utilizing comprehensive datasets from the Lusa News Agency and the STT Finnish News Agency, we've

expanded our models' linguistic scope, particularly incorporating Finnish to enhance cross-lingual performance and mitigate model overfitting risks.

Transitioning from CNNs to transformer-based models like BERT, we've embraced the capacity to generate deeper contextual embeddings, addressing the challenges of processing lengthy news articles through innovative methods. Our use of DistilUSE for creating multilingual sentence embeddings has significantly advanced our classification process, streamlining it by employing attention mechanisms for a more efficient, single-pass hierarchy prediction.

Adapting the AttentionXML model for multilingual contexts has improved our capability in extreme multi-label classification, integrating pre-trained multilingual embeddings and transformer models to boost contextual sensitivity and multilingual performance. Our novel approach to online multilingual news clustering, moving away from language-specific features and using DistilUSE for shared semantic space embeddings, simplifies the clustering process and enhances accuracy. This integration of dense and temporal features through a Rank-SVM model has fine-tuned our document clustering accuracy.

Task T2.5 News Summarization

In the second phase, we have substantially advanced automatic summarization, enhancing extractive and abstractive methods. We have focused mainly on improving abstractive summarization's factual consistency and linguistic depth across monolingual, cross-lingual, and speech content. Our efforts have concentrated on abstractive summarization, utilizing advanced neural approaches to overcome the shortcomings of extractive methods and ensure a refined content synthesis.

We have developed quality-aware abstractive summarizers by leveraging transformer architectures like BART and integrating new summarization metrics such as CTC scores. These models successfully re-rank summaries to align with human evaluations, significantly boosting summary quality and reliability. In addressing the scarcity of multilingual summarization resources, our work has led to methodologies that maintain semantic consistency across languages, ensuring content integrity across linguistic divides. This includes pivot-dependent

and independent strategies to enhance semantic similarity among summaries, overcoming traditional biases and inefficiencies.

Our explorations into speech summarization have resulted in an end-to-end system that directly converts speech to summarized text, leveraging models like Wav2Vec. This approach reduces error propagation from speech recognition and retains the speaker's nuanced intonation, which significantly advances comprehensive audio content summarization.

WP3 Joint Multilingual and User-Feedback Transfer Learning

Work package number	WP3	**Lead beneficiary:**	LIA
Work package title	Joint Multilingual and User-Feedback Transfer Learning		
Start month – End Month	M1 – M39		

WP leader: LIA

Participating Partners: DW, PRIB, IMCS, LIA, FhG

Task T3.1 Rich Transcription for Higher-Resourced Languages

Rich transcription means automatic transcription enriched by information like speaker labeling, gender detection, and can also include named entity recognition from speech.

In this task, we focus on high-resourced languages *i.e.,* languages for which a lot of audio/text paired training data is available. Recent advances in state-of-the art have been investigated, especially the use of self-supervised learning of speech representations (the wav2vec2.0 models), that takes benefit of audio data without text and that matches to the SELMA context: DW is able to provide thousands of hours of speech if no manual transcription is needed.

LIA is one of the main contributors to the LeBenchmark initiative. In 2021 and 2022 we trained several massive wav2vec 2.0 models for the French language and a paper has been published in the NeurIPS conference (NeurIPS 2021 Datasets and Benchmarks Track) that presents our

contributions, related to the SELMA project. These wav2vec 2.0 models are freely distributed to the community.

LIA also exploit this approach to develop its end-to-end ASR system and to build systems for different languages: French, English, Brazilian Portuguese, Modern Standard Arabic.

Fraunhofer also develop a similar approach and build systems at least for German, Spanish, and Russian languages. These models are now “production ready” (dockerized as webservices with API). Mixed language training for English and German was successfully applied to enhance the system’s capability to recognize anglicisms in German speech. Signal-augmentation strategies were also used during training to enhance the ASR-model's robustness towards low-quality speech (telephone, background-noise).

LIA also developed a system based on the same wav2vec2.0 neural architecture from named entity recognition and semantic information from speech. On this topic two papers were published – one for the LREC 2022 conference, and one paper for the SPECOM 2021 conference.

To better understand the behavior of the wav2vec2.0 models, the impact of the gender balance in the pre-training data to the final performance was investigated. LIA showed that in any case, it is better to start with gender balanced pretraining data. For instance, to process male (respectively female) speakers, we get better result if the wav2vec2.0 model has been pretrained on a gender balanced data than if this pretrained data contained only male (respectively female) speakers. LIA paper has been accepted to Interspeech 2022 and will be presented during the conference in September.

Ensuring proper punctuation and letter casing is a key post-processing step toward rich ASR transcriptions. This is especially significant for other textual sources where punctuation and casing are missing, like machine translation. Therefore, Fraunhofer jointly trained two token-level classifiers on top of a pre-trained BERT language model. It can restore both capitalization and punctuation marks (only ".?," for now) and is available in eight languages. In the second part of the project, both language span and punctuation marks will be increased.

Task T3.2 ASR for Low-Resourced Languages

The use of self-supervised learning (SSL) approaches is especially relevant when the availability of audio/text paired data is very low, since collecting audio without the transcription is easier. Even if audio recordings only are rare, it is possible to exploit SSL models pretrained on several different languages, even if the final target language was not present in the pretrained data.

So, the work we made for Task T3.1 was also useful for Task T3.2, and in addition to the use of monolingual wav2vec2.0 model, we also investigated the use of multilingual models, like the XLSR-53 wav2vec2.0 model released by Meta AI.

LIA simulated a low resource scenario for French language in the LeBenchmark NeurIPS paper mentioned in Task 3.1 and could confirm the strong improvement provided by such an approach. LIA also build an ASR system for Tunisian that got the best Word Error Rate in the IWSLT 2022 campaign we attended in 2022. Apart from that Fraunhofer worked on Russian language using the fine-tuning strategy over multilingual wav2vec2.0 models.

We also investigated the construction of an end-to-end speech-to-text named entity recognition system when no speech data are annotated with named entity. We proposed an approach that allows us to inject textual information in a neural network fed by speech only. This approach is described in an Interspeech 2022 paper that has been accepted and will be presented in September.

Task T3.3 Text and Speech Machine Translation

For this task, LIA has been focusing on the production of direct speech machine translation models that leverage pre-trained blocks for speech called wav2vec 2.0. By producing models that translate speech directly into the targeted language, without the production of transcriptions, we can produce language resources for low-resource languages that lack written form, and/or for which not many resources are available. Directly translating from speech also has other advantages: the speech can provide clues for vocabulary and speaker disambiguation that are sometimes lost in transcription. To benchmark the SELMA technology to state-of-the-art systems, this year LIA participated and helped organize the IWSLT 2022 (International Workshop on Spoken Language Translation). For the low-resource speech translation task (a

task LIA organized and participated in) we trained direct speech machine translation models that translated Tamasheq speech into French text using only 17 hours of parallel data. We also trained wav2vec 2.0 models, the pre-trained blocks we can re-use for different speech tasks, in Tamasheq and close languages. All the pre-trained models are freely available at HuggingFace, and the recipe for our best setup was submitted as a recipe in the SpeechBrain library. As future directions for more effective direct speech translation models in low-resource settings, we intend to investigate a deeper integration between pre-training and fine-tuning steps, and the levering of multilingual information.

As written above, LIA is also one of the main contributors to the LeBenchmark initiative and we trained several massive wav2vec 2.0 models for the French language. In these settings, these pre-trained speech blocks were used for training direct speech translation models for translating speech in French into English text.

Task T3.4 Automatic Post-Editing

Task T3.5 Voice Conversion Synthesis

In this section, the neural network-based architecture developed during the first year of the SELMA project is presented, in addition to the update for this delivery.

During the first year of the SELMA project, the first version of the TTS engine was released. The system is using an end-to-end model based on VITS [Kim et al., 2021] architecture. To train the speech synthesis engine, we use the audio news bulletins that are produced by DW's Brazil department. The audio files have been downloaded from YouTube and the scripts were retrieved from GitHub in a repository with all the text scripts that DW uses to produce their weekday news podcasts.

For this delivery, we include new data collected until now to improve the performance of the text-to-speech engine; this represents a 30% improvement in the amount of data. With this update, we are trying to address the issues that were raised during the consortium meeting in Avignon. To do this, we include an additional data cleaning stage to filter samples that contain

errors in the alignment or the segmentation, then we use a diarization system to remove silence and music in the training data.

WP4 Platform Integration

Work package number	WP4	**Lead beneficiary:**	IMCS
Work package title	Platform Integration		
Start month – End Month	M1 – M39		

WP leader: IMCS

Participating Partners: DW, PRIB, IMCS, LIA, FhG

Task T4.1 Integration of NLP Components

This task has been the prime focus of WP4 activities from the very start of the SELMA project (40 internal meetings with minutes, Docker development lab for x86 and ARM architectures set up) and is extensively covered in the deliverables D4.1 “Platform architecture and API documentation”, D4.2 “Initial platform release with the primary NLP pipeline”, D4.3 “Intermediate platform with continuous massive stream learning NLP capabilities”, and D4.4 “Final platform release with full continuous massive stream learning capabilities”. Integration of NLP components happens in the SELMA NLP pipeline consisting of the three backend core components:

1. **Maestro-Orchestrator** for NLP job queueing according to dependency DAG,
2. **Token-Queue** for scheduled access to shared NLP worker pool,
3. **Docker-Spaces** for massive distributed on-demand scaling of the NLP worker pool.

To freely test these three SELMA NLP backend components, in deliverable D4.1 “Platform Architecture and API Documentation” was introduced the **SELMA NLP Service Testing, Configuration, and Orchestration Use Case 0 (UC0)** which through multiple iterations eventually transformed into SELMA OSS (Open Source Software) described in D4.4 “Final

platform release with full continuous massive stream learning capabilities". **UC0 SELMA OSS was** not envisioned in the original project proposal and is accessible at address https://selma-project.github.io/. This enabled efficient testing and integration of NLP components developed within the SELMA project prior to their embedding into the limited-access commercial platforms for primary Use Cases UC1 (media monitoring) and UC2 (news production). Scalability of these approaches on various x86 and ARM GPU architectures as well as cloud infrastructures is described in the D4.4 along with the description of the SELMA OSS release.

Task T4.2 Integration of Continuous Massive Stream-Learning Components

Having established the core NLP processing platform in Task 4.1, here we focused on the logical integration of the various NLP components developed within the SELMA WP2 "Continuous Massive Stream Learning" and WP3 "Joint Multilingual and User-Feedback Transfer Learning". The key logical integration problem of the NLP components is converting the output of one NLP component to the input format expected by the next NLP component in the NLP pipeline DAG (Directed Acyclic Graph): traditionally various NLP modules use different JSON input/output schemas not compatible with other NLP modules in the pipeline. To deal with JSON schema conversion "on-the-fly" between the various NLP modules Maestro-Orchestrator DAG engine supports two approaches: (1) converting all inputs and outputs to/from the shared legacy "SUMMA JSON" schema or (2) execution of the dynamically supplied JavaScript JSON conversion script. The second (2) JavaScript based approach is implemented also in the initial "white" version of the Testing and Configuration Use Case 0. In the second half of the SELMA project following UC0 "yellow", "green" and "red" versions adopted a novel DockerSpaces enabled approach to NLP component integration, where the entire NLP pipeline is executed in the JavaScript in the frontend rather than being sent to the Maestro-Orchestrator backend for execution. The Continuous Massive Stream Learning components being integrated into the UC0, UC1, UC2 are described in the deliverables of WP2 and WP3 along with the overall User Experience covered in the deliverables of WP1.

Task T4.3 UI/UX for the Multilingual Media Monitoring Use Case

The Graphical User Interface for the Multilingual Media Monitoring Use Case (UC1) was developed as part of the Monitio platform. Priberam accommodated the UX/UI requirements

set forth in the SELMA WP1. In SELMA, we are leveraging the efforts from the commercial Monitio platform, allowing us to focus on the NLP research aspects of the Media Monitoring problem, mainly the activities related to Natural Language Processing, Transfer Learning, User-Feedback Learning, Stream Learning, and the activities related to performance scalability for processing massive streams.

Task T4.4 UI/UX for the Multilingual News Production Use Case

The Graphical User Interface for the Multilingual News Production Use Case (UC2) was developed as part of the plain X commercial media monitoring platform. During the SELMA project plain X platform UI/UX has been completely redesigned to meet the requirements set forth in the SELMA WP1. The f redesigned UI/UX is described in the deliverables of WP1.

WP5 Evaluation

Work package number	WP5	**Lead beneficiary:**	DW
Work package title	Evaluation		
Start month – End Month	M4 – M39		

WP leader: Deutsche Welle

Participating Partners: DW, PRIB, IMCS, LIA, FhG

Task T5.1 Technical Evaluation

A detailed overview of all components and technologies covered in the project was established, with the type of evaluation specific to each component, in order to be able to track progress and evaluation. A technical evaluation was performed by the technology partners on the individual components developed in the project, as well as on the integrated platforms. Details are covered in the individual technical work packages.

Task T5.2 User Evaluation

In P1, we established the evaluation plan, with the objectives, the evaluation methodology, and a detailed overview of planned technical and user evaluation. We created a table of envisaged evaluations at the different levels and per partner, serving as a work sheet to track and plan evaluations throughout the project.

User Evaluation took place on all three use cases, where we assessed the overall performance of the platform prototypes from a user point of view, i.e., plain X for the news creation use case, Monitio for the monitoring use case, and SELMA OSS for the integration and orchestration use case. The focus of specific user applications was on audio podcast creation.

We performed user evaluation on specific components that were ready for such assessment, including ASR, speech translation, TTS, Named Entity Recognition through prototypes/UIs provided by the technology developer.

In P2, technical evaluation focused on the final modules and platforms of all technologies developed in SELMA, including ASR enhancement, diarization and speaker recognition, speech-to-translated-text, speech synthesis, NER/NEL, summarization, integration of demonstrators and orchestration.

User assessment efforts targeted NER analysis, speech-to-translated-text, usability evaluation of the three primary demonstrators plain X, Monitio and the OSS, and the use case applications on speaker, summarization, podcasting and diversity. Special focus was the further development of a user-friendly benchmarking tool for transcription, translation and voice-over engines,

WP6 Impact

Work package number	WP6	**Lead beneficiary:**	DW
Work package title	Impact		

Start month – End Month	M1 – M39

WP leader: Deutsche Welle

Participating Partners: DW, PRIB, IMCS, LIA, FhG

Task T6.1 Dissemination

This task was about establishing the dissemination strategy and conducting awareness activities through the website and other (social) media channels. Major achievements of the task include the creation and the maintenance of the website www.selma-project.eu which was active from Month 3 on and attracted more than 30.000 views. We set up the communication and dissemination strategy (for more details see D6.2 Impact Plan). All partners contributed to this task by providing input and material to populate the website and made it an interesting place for getting to know the objectives and results of the SELMA project. A dissemination kit containing a set of key visuals of the SELMA project, such as a flyer, poster, and banner has been created and maintained as well as a promotional video and prototype walkthroughs. The kit was accordingly adapted to COVID19-influenced requirements of mostly online events (e.g., through creation of video call background images) in the first period of the project, where partners have used it at 22 academic and industry dissemination events and for SELMA's communication channels. In the second period we went back to mainly physical meetings, where we used the toolkit as anticipated with flyers and roll-ups. In total we were actively participating in more than 60 events and published more than 30 papers (for more details see D6.4 Interim Impact Report and especially D6.6. Final Impact Report).

Task T6.2 Exploitation

This task was about how SELMA output can be exploited by the partners themselves and others. A central focus was how the Open-Source platform and the NLP components including the two media monitoring platform (Monitio) and the media production platform (plain X) can be used and exploited for the public and for media companies for analytics and journalistic purposes. A first set of exploitation objectives and opportunities had been laid out in D6.2 Impact Plan in the first period. In the second part we made an assessment of business opportunities, especially

for the platforms and plain X was evolving into a product. Consequently, a product website was developed and launched (www.plainx.com) and first negotiations with clients were led. In parallel, plain X was rolled out at DW as a tool for translation, but also for creating subtitles to meet and adhere to accessibility standards. The project participated in the EU Innovation Radar with three submissions; all of them were selected and promoted on the corresponding EU website. Major activities were the establishment of an IPR framework.

Task T6.3 Data Management

Since SELMA involved extensive work with language data, this task identified what kind of data was collected, processed and generated by the SELMA consortium, what personal data is collected by the Use Case 1 and Use Case 2 platforms and what data protection means are applied, which datasets are IPR-protected and how they are handled. Additionally, this task identified which language datasets created within SELMA may be released in the public domain (considering IPR and GDPR aspects) and which datasets (majority) are for internal use only by the SELMA consortium.

A major change in the data management plan was made regarding Diversity Use Case (as a part of Use Case 1): to prevent the risks of potential personal data infringement, we abandoned the idea (and its prototype implementation) of using sensitive Wikidata properties (like education, nationality, sexual orientation) of named entities; only non-sensitive Wikidata properties, such as binary gender and age, are collected and stored to support the diversity use case.

The most notable datasets that were created within SELMA for the internal development of NLP components are:

1. a large collection of audio/video recordings (with metadata) for 19+ DW languages – more than 16,000 hours provided by DW and used for pre-training the SELMA Foundational Multilingual Wav2Vec 2.0 model (open source);
2. a dataset of transcribed audio news bulletins for less-resourced DW languages – Brazilian Portuguese (87 hours, several speakers), Urdu (10 hours), Amharic (10 hours) and Bengali (5 hours) provided by DW and used for training quality TTS models for generating podcasts and voice-overs (Use Case 2);

3. a dataset of hierarchically annotated named entities mentioned in news articles of selected languages – Latvian (740 articles), Ukrainian (300 articles), Russian (160 articles), Turkish (100 articles) and Dutch (50 articles) provided and annotated by DW and IMCS and used for training (extending) a multilingual named entity recognition (NER) model which is integrated into Use Case 1.

Task T6.4 Communication

This task was about the communication of the project and its results to a wider audience (beyond the consortium's own community and stakeholders). It includes European Commission research groups and collaboration events and the public at large.

Major achievements were the set-up and the maintenance of a blog section within the homepage, mainly aimed at the general public. All partners have created articles / blog posts around human language technologies and artificial intelligence related to their specific knowledge in the domain. In total 14 of those articles were published. The contributions were well received from the website audience. The project has set up its own user group (involving media monitoring and production stakeholders) and conducted two user group meetings to gather feedback. SELMA participated in and contributed to several BDVA activities (online workshops and conferences in Sweden and Spain).

WP7 Management

Work package number	WP7	**Lead beneficiary:**	DW
Work package title	Management		
Start month – End Month	M1 – M39		

WP leader: Deutsche Welle

Participating Partners: DW, PRIB, IMCS, LIA, FhG

Task T7.1 Project Administration and Resource Monitoring

This task focused on establishing the project management tools and procedures, communication means and mechanisms to ensure smooth collaboration. A mailing list was established at the start of the project together with a shared space (MS SharePoint for working jointly on documents, Atlassian Confluence for archiving files).

To manage the project administratively and technically, bi-weekly general calls on MS Teams were held. Additionally, each work package organized its own calls or meetings whenever required.

Due to the pandemic, no face-to-face meetings could be held in the first 17 months of the project. Instead, these three virtual consortium meetings were organized by the coordinator DW:

1. Kick-off meeting (11, 12, 14 January 2021)
2. 2nd consortium meeting (13,14,16 September 2021)
3. 3rd consortium meeting (8-10 February 2022).

The first physical meeting (4th consortium meeting) took place at LIA's premises in Avignon, France, on 7/8 June 2022.

In the second part of the project the following face-to-face meetings took place:

1. 5th Consortium Meeting and First User Day, 11-13 October 2022, Bonn, Germany
2. 6th Consortium Meeting, 18/19 April 2023, Riga, Latvia
3. 7th Consortium Meeting, 19/20 September 2023, Lisbon, Portugal
4. 8th Consortium Meeting and Second User Day, 14-16 November 2023, Avignon, France

As required by the Commission, the internal review (13 September 2022) was an online event.

Task T7.2 Quality Control and Work Plan Monitoring

A common process for the preparation and quality control of project deliverables was established. Each deliverable has been reviewed by a partner that has not been involved – or at least not strongly involved - in the writing of it and by the Project Coordinator prior to the submission. A deliverable template was created by the coordinator and was used for all deliverables.

Work on risk management started early to identify potential threats to the success of the project. Measures on how to minimize these risks and to mitigate their impact if required have been laid down. No significant risks to the success of the project were identified until the end of the project.

The deliverable associated to this task D7.1 Quality Assurance and Risk Assessment Plan was submitted (M6). It provides guidelines about the general project organization, information management, reporting and quality assessment procedures as well as risk management.

The Interim Periodic Progress Report (D7.2) was mistakenly scheduled for M18. As the PPR was due in late August 2022 (60 days after the completion of the reporting period M1-M18), the associated deliverable was submitted on 16 September 2022.

Due to the complexity of the project challenges, the consortium asked for a three-month project extension, which was granted by the Commission in the required contract amendment in early November 2023.

Task T7.3 Communicating with and Reporting to the EC

This task was about prompt communication and reporting to the EC, including the EC's project officer. A good working mode has been established and the midterm progress report was submitted providing a management-level overview of project activities carried out in the first project period. It contains a description of the overall scientific, technical, and innovative objectives and a progress report with respect to milestones and deliverables of the project as well as the financial statements.

WP8 Ethics Requirements

Work package number	WP8	**Lead beneficiary:**	DW
Work package title	Ethics Requirements		
Start month – End Month	M1 – M39		

WP leader: Deutsche Welle

Participating Partners: DW, PRIB, IMCS, LIA, FhG

The SELMA ethics deliverable was submitted early in the project (M3). The ethics process grouped the ethical issues which arise into six broad categories: Protection of personal data, Copyright protection, Ethical implications of SELMA technologies, General ethical concerns related to open-source release of novel analytics technologies, The social impact of automation and Sex and gender balance. These aspects, and the project's response to them, were discussed in D8.1 Ethics Deliverable. As a result of the first review two ethics advisors were consulted which handed in two reports (in March 2023 and in March 2024) which were acted upon. The advice of the first report for example led to a reshape of the Diversity Application. Both ethics reports were added to D8.1 which was updated in M27 and M39.

Ethics considerations are also part of the data management, project management and evaluation reports. A DPIA for the project as a hole was conducted and Priberam conducted a DPIA for the Monitio platform.

Impact

The information in section 2.1 of the DoA is still relevant, and no update is needed.

Access provisions to Research Infrastructures

Not applicable.

Resources used to provide access to Research Infrastructures

Not applicable.

Updates of the plan for exploitation and dissemination of result (if applicable)

Not applicable.

Update of the data management plan (if applicable)

The initial DMP was described in D6.1 (M6) and updated in D6.3 (M18) and in D6.5 (M39). There are no further updates to report in D7.2.

Follow-up of recommendations and comments from previous reviews (if applicable)

During the interim review a potential issue in respect to ethics and the use of personal data was raised. As a result the project has reached out to two external advisors (more details in the updated D8.1 Ethics Deliverable).

Deviations from Annex 1 and Annex 2

Tasks

All tasks have been progressed as scheduled.

Use of resources

For the use of resources please see: SELMA-Technical-Report-PartB-P1.pdf.