Research and Innovation Action (RIA) H2020 – 957017

SELMA

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D7.2 Interim 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	0.9
Contractual Date	16 September 2022
Delivery Date	16 September 2022
Dissemination Level	PU

Version History

Version	Date	Description
0.1	19.07.2022	First draft
0.2	01.08.2022	Updated draft
0.3	15.08.2022	Integrated contributions from all partners
0.9	06.09.2022	Draft version
1.0	16.09.2022	Final version for submission

Table of Contents

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

Large amounts of multilingual information in the form of data is all around us and growing strongly. Still, the potential to fully take advantage of these digital content streams based on machine learning has remained widely untapped.

SELMA tackles these potentials from two sides: by advancing language technologies from a research perspective and by integrating concrete technological improvements into a platform which to a large extent will be available open-source for the public and the (media) industry.

A focus of the SELMA work lies in a unified approach to multilingual media monitoring and content production by leveraging and contributing to advances in deep learning, in particular in language modelling, 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, an international broadcaster) significant progress into “shaping speech and text technologies for media monitoring & the newsroom” has already been made in the first period of the project.

SELMA is organized around 2+1 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 example) and Press Agency Analysis.

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 and translation. It comprises 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 is 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 period of the project.

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

A core objective of SELMA is to process extreme large amounts of audiovisual and textual data. 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 and the open-source platform being capable of processing up to 10 million items per day (although scalability tests are scheduled in the second period of the project) was constructed. See deliverable D4.2 Initial platform release with the primary NLP pipeline for details.

Related milestones:

MS2: Release of Architecture Agreement (M10)
MS4: Release of Platform Release v1(M15)

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

This objective concerns 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 this context 17 academic papers have been 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.

Related milestones:

MS3: Release of Initial Prototypes (M12)

Objective 3: Knowledge transfer across tasks and languages

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

tasks. Initial progress of transfer learning has been made and described in the technical deliverables (D2.2, D2.3 and D3.3), but will be an ongoing activity in the second period of the project.

Objective 4: Enable media monitoring analytics for decision-making

This objective is 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. As detailed in D1.2, SELMA 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.

Related milestones:

MS1: Release of Requirements Analysis (M6)
MS5: Release of User Evaluation 1 (M15)

Objective 5: Enable multilingual content production workflow

This objective is to provide a content production workflow by leveraging multilingual transcription and translation models trained within SELMA. The target group are journalists and editorial production teams. The plain X user interface has been developed to meet 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 has started on the Podcast Producer tool, a separate application whose purpose it is to use UC0's speech synthesis modules to support the semi-automated creation of news podcasts.

Related milestones:

MS1: Release of Requirements Analysis (M6)
MS5: Release of User Evaluation 1 (M15)

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

This objective aims to improve and fine-tune deep learning models from user feedback based on novel deep learning methods. As a precondition an end-to-end model for named entity recognition from speech without paired training data has been built in the framework of the SELMA project. This will enable the project to investigate novel ways to generate massive amounts of training data for the post-editing task. See deliverable D3.3 Initial release of post-editing and user feedback capabilities for details. In the second period of the project SELMA will ingest annotations from users gained in an editorial news production setting into the learning models and thus 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. With the set-up of the SELMA open-source platform many technological developments within SELMA are 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. Still, major activities as well as the related milestone (MS16 Sustainability Plan) are scheduled for the second period of the project.

Objective 8: Dissemination and communication of the SELMA project outputs

This objective mainly concerns work done in WP6. SELMA has set up a user group and engages with all relevant stakeholders in the language technology and innovation chain, including broadcasters, commercial players, EU agencies and the relevant research communities. The milestones relating to this objective are scheduled for the second period of the project (MS6 User Day 1, MS11 User Day 2). However, measurable progress has been made through various project presentations and engagement in external events as well as several project publications. See deliverable D6.2 Impact Plan for details.

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 – M36		

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 is being applied into three main use cases, Multilingual Media Monitoring (UC1), Multilingual News Content Production (UC2) and SELMA OSS 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.

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.

Task T1.3 Multilingual Media Monitoring Prototype

Within Use Case 1 (UC1), we have been integrating 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 D2.1 for technical details), while others are changes to the platform to allow learning from user feedback. Another prototyping effort which is underway is integrating the SELMA NLP Service Orchestration (UC0) as *Monitio*'s job orchestrator, which will allow *Monitio* to scale (see D4.1 and D4.2). As detailed in D1.2, we have so far integrated SELMA results into Monitio from the following WP2/3 components: Multilingual News Clustering, Multilingual Topic Detection, and Rule-Based Entity Correction. More recently since the writing of that deliverable we have also integrated a new version of the Named Entity Recognition component trained on the dataset produced by SELMA in T2.2. Regarding the Advanced Content Analysis scenario goals, we have also set up a Wikidata processing pipeline to retrieve relevant properties of known entities (such as gender, age, and occupation), which was a prerequisite for further developments of this use-case scenario.

Task T1.4 Multilingual News Production Prototype

Within Use Case 2 (UC2), we are integrating 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* is using the SELMA orchestration to schedule and execute NLP jobs (see D4.1 and D4.2). 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) but also from many cloud providers (Azure, Google, etc). The *plain X* user interface is being developed to meet SELMA's UC2 requirements (see D1.2). We are also integrating in the prototype results from WP2/3 into *plain X,* Automatic Speech Recognition, Speech Translation and Punctuation modules. At the present date, we've integrated LIA's TTS Model: 9 PT-BR voices trained by DW speakers (T3.5), LIA's Wav2Vec ASR for French (T3.1), LIA's Speech Translation ASR for French->English (T3.3) and the SELMA Orchestration Core (WP4).

Work has started on the Podcast Producer tool, a separate application whose purpose it is to use UC0's speech synthesis modules to support the semi-automated creation of news podcasts.

WP2 Continuous Massive Stream Learning

Work package number	WP2	**Lead beneficiary:**	Fraunhofer
Work package title	Continuous Massive Stream Learning		
Start month – End Month	M1 – M36		

WP leader: Fraunhofer

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

Task T2.1 Cross-lingual Stream Representations

This task focuses on learning contextual word and entity representations captured from a live news article stream. However, the extensive data scale makes this task particularly challenging, in addition to the emphasis on serving across several languages simultaneously. Hence, to enable knowledge transfer from higher- to lower-resourced languages, we aim to learn a cross-lingual representation space, i.e., a representation where word contexts from different languages map into a shared space, to enable knowledge transfer from higher- to lower-resourced languages. To this end, Priberam has successfully trained multilingual entity embedding representations for about 15 million entities extracted from Wikidata, which has a Wikipedia page; at this point, we can train models based on 40 Wikipedia languages. Our method follows Ganea and Hofmann [2017]; we bootstrap the representations from BPE multilingual word embedding and use a max-margin objective from a positive word-entity co-occurrence on Wikipedia and a negative distribution. We evaluate the quality of the models by computing the similarity with the 100 nearest neighbors expecting that the related entities rank first. Our evaluation shows that the multilingual embeddings have better on our test set than the English monolingual ones. We use and evaluate these representations on the downstream task of Entity Linking. The proposed method will allow incrementally training new entities from the stream or from Wikidata/Wikipedia. Priberam developed a complete pipeline that goes from the downloading of the latest Wikipedia and Wikidata to the training of the representations that automates the whole process of creating a database with the

representations together with the needed entity properties to be used on the deployments, allowing periodically updates with the most recent entities from Wikidata.

Task T2.2 Named Entity Recognition and Linking

This task aims to develop statistical models for detecting entity mentions within news article streams and learning a mapping of these entities to a knowledge base. This step is fundamental to performing content enrichment on the data stream and the foundation for the post-editing task.

Even though entity mentions are inherently nested and hierarchical, most of the available datasets for the task are annotated using simple flat mentions. Priberam, IMCS, and DW have annotated a multilingual nested hierarchical NER dataset based on a common ontology. We already have annotated more than 15000 documents in Portuguese, Spanish, French, English, German, Latvian and Dutch. DW is currently annotating Urdu, Arabic, Turkish and Arabic, and IMCS is annotating Russian and Ukrainian.

Priberam developed two entity recognition models, the stack-LSTM and the Bi-affine model, for hierarchically nested entity recognition. Both models presented similar performances in our new test sets, where the stack-LSTM has shown better performance characteristics to be deployed in the UC1 scenario. In a first approach, we trained Spanish, English, Portuguese, German, French, and Latvian standalone models to be deployed on UC1. We are investigating the behavior of the models when jointly training one single model for all the languages and evaluating the zero-shot performance on languages unseen in the training data. Preliminary results show that results improve on the test sets for the training languages and that the zero-shot languages have a very promising performance allowing to reduce the effort on the annotation task.

Based on the representations developed for task T2.1, Priberam developed a multilingual Named Entity Linking system capable of disambiguating entities against a Knowledge base of about 15 million entities. Our approach follows Dynamic Context Augmentation (DCA) (Yang et. al. 2019) with improved multilingual word and entity base representations together with an end-to-end mention type classification and better coherence model. Our model

evaluates at the state-of-the-art for English in the CoNLL-Yago dataset and shows good multilingual capabilities when evaluated on the TAC and Voxel datasets exhibiting outstanding speed and memory footprints that enable us to deploy on UC1.

Priberam is now researching using this system in the stream learning and NIL clustering of new and trending entities.

Task T2.3 Story Segmentation

This task mainly aims to segment long audio segments into meaningful units, providing speaker clustering, speaker recognition, and topic segmentation. For speaker clustering, the identity of the speakers is unknown, and the system provides only labels for segments of the same speaker appearing multiple times in one file. In the first period of the project, Fraunhofer focused on text-independent speaker diarization and recognition systems when the identity of the speaker is based on how the speech is spoken, not necessarily on what is being said. Fraunhofer followed ECAPA-TDNN architecture which can eliminate some limitations of the x-vector embeddings. This new model extends the temporal attention mechanism even further to the channel dimension and enables the network to focus more on speaker characteristics that do not activate on identical or similar time instances. This approach can segment speakers in spoken content with high performance regardless of language context. Therefore, these modules are developed in a language-independent fashion with the help of multilingual speaker embeddings. Fraunhofer is now dealing with the combination of news classification and speech recognition modules together with speaker segmentation to classify news stories inside the spoken content.

Task T2.4 Online News Classification and Clustering

Text classification and clustering consist of aggregating similar news stories from the data stream and classifying documents with topic labels from a particular set.

Priberam approached the classification task where we mainly used IPTC subject codes (it covers 1404 topic labels distributed over a hierarchy of three layers), having label names and descriptions in seven languages (English, German, French, Portuguese, Spanish, Italian, and Japanese). For the classification task, we first released a new multilingual model for the IPTC

topic classification based on the AttentionXML model that improves performances in the test sets for Portuguese, English, and Spanish compared to the previous SUMMA model. An additional benefit of this model is that it enables to backtrack the attention to the input words enabling model explainability. This model has been deployed on Monitio, and it is being used for the 30 DW languages, among others.

On the clustering task Priberam has developed a new model, presented at the SIIG conference on the TextToStory workshop. We empirically demonstrate that the use of multilingual contextual embeddings as the document representation significantly improves clustering quality. We challenge previous cross-lingual approaches by removing the precondition of building monolingual clusters. We model the clustering process as a set of linear classifiers to aggregate similar documents and correct closely related multilingual clusters through merging in an online fashion. Our system achieves state-of-the-art results on a multilingual news stream clustering dataset, and we introduce a new evaluation for zero-shot news clustering in multiple languages. We made our code available as open-source.

Our latest model is capable of processing 3.2 documents per second while maintaining a cluster pool of 25,000 clusters while at the same time improving the state-of-the-art on multilingual evaluation dataset by 5% and achieving impressive results on the extended dataset for unseen language during training. This model is currently deployed on UC1 (Monitio).

Task T2.5 News Summarization

This task focuses on summarizing the information conveyed by a longer source document using state-of-the-art neural methods. The approaches for automatic summarization can be divided into two categories: extractive and abstractive methods. Our target is the abstractive summarization methods where we want to process the summary by generating new text that paraphrases the most relevant parts of the source document. We first investigated transformer-based sequence-to-sequence architectures where our main goal is to improve the factual consistency and the overall quality of the generated summaries. We leverage recent advances in summarization metrics to create quality-aware abstractive summarizers. Priberam proposed an energy-based model that learns to re-rank summaries according to one or a combination of

different summarization metrics. This model consists of a BERT that receives the document and a candidate summary and is finetuned to output a score that mimics the chosen metrics. The results show that sampling candidates from an abstractive system (BART, PEGASUS, …) and re-ranking them with this model consistently improves the quality of the produced summaries over the usual beam search, achieving over 1% improvement in ROUGE scores and up to 3% improvement in automatic factual consistency metrics on benchmark datasets for abstractive summarization.

Current research focuses on cross-lingual end-to-end summarization, where documents from a specific language are summarized in a different target language. Leveraging LIA's speech knowledge, Priberam is also researching end-to-end Speech 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 – M36		

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

We recently started working on this task that aims to correct some mistakes made by the speech and language processing tools.

For speech processing, a central idea is to take benefit of previous manual corrections made by the users. The approach we proposed for task 3.2 that enabled to inject textual information into an end-to-end speech-to-text system has also been designed for this purpose. Experiments are in progress. We also studied the literature and plan to investigate the use of a translation memory (the translations validated by the users) conjointly to the use of our neural approaches. This approach recently shows promising results.

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 – M36		

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 (29 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” and D4.2 “Initial platform release with the primary NLP pipeline”. 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, we have introduced an open-source **SELMA NLP Service Orchestration Use Case 0 (UC0 SELMA OSS)** not envisioned in the original project proposal and accessible at address https://selma-project.github.io/. This enables 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 will be the focus in the second half of the project.

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

Having established the core NLP processing platform in Task 4.1, here we focus 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 Testing and Configuration Use Case 0 and we intend to strengthen this approach throughout the second half of the SELMA project. The Continuous Massive Stream Learning components being integrated into the UC0, UC1, UC2 are described in the deliverables D2.2 “Initial release of stream learning and entity linking capabilities”, D2.3 “Initial release of segmentation, summarization and news classification

capabilities", D3.2 "Initial release of transcription, punctuation, translation, voice synthesis capabilities", and D3.3 "Initial release of post-editing and user feedback capabilities".

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

The Graphical User Interface for the Multilingual Media Monitoring Use Case (UC1) is being developed as part of the Monitio platform. Priberam accommodated the UX/UI requirements set forth in the SELMA WP1 and more specifically in the deliverable D1.1 "Use Case Description and Requirements". The first version of the redesigned UI/UX is described in the deliverable D1.2 "Initial Prototype Report". 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 and Stream Learning and also 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) is being 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 and more specifically in the deliverable D1.1 "Use Case Description and Requirements". The first version of the redesigned UI/UX is described in the deliverable D1.2 "Initial Prototype Report".

WP5 Evaluation

Work package number	WP5	**Lead beneficiary:**	DW
Work package title	Evaluation		
Start month – End Month	M4 – M36		

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

Task T5.2 User Evaluation

We have established the evaluation plan, with the objectives, the evaluation methodology, and a detailed overview of planned technical and user evaluation. We have 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 has taken place on all three use cases, where we assess 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 has been on audio podcast creation.

We have 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.

WP6 Impact

Work package number	WP6	**Lead beneficiary:**	DW
Work package title	Impact		
Start month – End Month	M1 – M36		

WP leader: Deutsche Welle

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

Task T6.1 Dissemination

This task is 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 of the website www.selma-project.eu which was active from Month 3 on, and the set-up of 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 make it an interesting place for getting to know the objectives and first 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 as well as a promotional video. The kit was accordingly adapted to COVID19-influenced requirements of mostly online events (e.g., through creation of video call background images). It has been used by partners at 22 academic and industry dissemination events and for SELMA's communication channels. Since the beginning of the project, 13 publications have been published; an additional six are in review mode.

Task T6.2 Exploitation

This task is about how SELMA outputs can be exploited by the partners themselves and others. A central focus will be how the Open-Source platform and the NLP components 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 have been laid out (for more details see D6.2 Impact Plan). Results of WP2, WP3 and WP4 have been integrated into the platforms. Major activities in this task are timetabled for the second half of the project which includes the study of the marketplace (State of the Art), the assessment of business opportunities, the establishment of an IP framework and finally the development of an exploitation agreement in the third year.

Task T6.3 Data Management

As the SELMA project involves extensive work with language data (textual, audio and video, including metadata), this task identifies what kind of data is being collected, processed and generated by SELMA, which datasets are IPR-protected and how they are handled, what

personal data is collected by the Use Case 1 and Use Case 2 platforms and what data protection means are applied, which datasets and input streams potentially contain random personal data and what measures are applied to exclude or minimize inclusion of such data. Additionally, this task identifies which training or test datasets created within SELMA may be released as open data (considering IPR and GDPR aspects) and which datasets are for internal use only by the SELMA consortium.

For the media monitoring use case, the SELMA platform is currently ingesting ~300,000 news articles per day from more than 7,000 target sites covering almost completely Portugal and Spain, and the main media sites for other geographies in Europe, Latin America, and Oceania. For training speech recognition and speech summarization models, DW has collected and provided to the SELMA technology partners audio/video recordings (with metadata) for 16 languages - more than 11,000 hours in total. For training a text-to-speech model for Brazilian Portuguese, a dataset of transcribed audio news bulletins of nearly 90 hours was also provided by DW. For training named entity recognition (NER) and linking models, in addition to prior datasets created by Priberam, DW is currently annotating ~500 Arabic and ~500 Turkish news articles, while IMCS has annotated ~750 Latvian news articles and is currently validating automatically pre-annotated sets of ~500 Russian and ~500 Ukrainian news articles. We plan to release these additional NER datasets as SELMA open data.

Task T6.4 Communication

In this task is 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 are the set-up of a blog section within the homepage, mainly aimed at the general public. All partners have created articles around human language technologies and artificial intelligence related to their specific knowledge in the domain. The contributions are well received from the website audience according to high view numbers. The project has set up its own user group (involving media monitoring and production stakeholders) and

conducted a first user group meeting. SELMA participates and contributes to BDVA activities.

WP7 Management

Work package number	WP7	**Lead beneficiary:**	DW
Work package title	Management		
Start month – End Month	M1 – M36		

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, Confluence for archiving files).

To manage the project administratively and technically, bi-weekly general calls on MS Teams were established. Additionally, each work package organizes 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.

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 is reviewed by a partner that is not 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 is 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. Risk management activities will be continued 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.

Task T7.3 Communicating with and Reporting to the EC

This task is 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 is on its way providing a management-level overview of project activities carried out in the first 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 – M36		

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 issues, and the projects response to them, were discussed in D8.1 Ethics Deliverable. They are and will be part of the data management, project management and evaluation reports.

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). There are no further updates to report in D7.2.

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

Not applicable.

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.