Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D5.1 Evaluation Plan

Work Package	5
Responsible Partner	Deutsche Welle
Author(s)	Olga Kisselmann, Peggy van der Kreeft
Contributors	Yannick Estève, Guntis Barzdins, Afonso Mendes, Christoph Andreas Schmidt, Tugtekin Turan
Reviewer	Guntis Barzdins
Version	V1.0
Contractual Date	31 December 2021
Delivery Date	31 December 2021
Dissemination Level	Public

Version History

Version	Date	Description
0.1	17/10/2021	Initial Table of Contents (ToC)
0.2	1/12/2021	Main input from partners
0.3	8/12/2021	Components List added
0.4	20/12/2021	Internal Review version
0.5	23/12/2021	Finalization
1.0	29/12/2021	Publishable version

Executive Summary

Evaluation is central to SELMA's activities. It takes the ambitious technological research and prototype development to a next level by determining its added value, its strengths and weaknesses and room for improvement, and - last but not least - its usefulness for the media world, our focus user group, for the two use cases, multilingual media monitoring and news production.

This document provides an overview of the initial evaluation plan for the SELMA platform and its use cases. It showcases the evaluation activities planned and methods involved, including functionality and component testing, evaluation activities involving specially developed demonstrators as well as the integrated platform, user evaluation of the use cases, and the process and logic behind selection and recruitment of the test users. The emphasis in all activities will be put on extensive qualitative and quantitative technical testing as well as user-centered testing. The document at hand is the first iteration of three reports. It will be followed by D5.2, the Interim Evaluation Report in Month 24, and D5.3, the Final Evaluation Report at the end of the project in Month 36.

This report presents the evaluation objectives in section 2, as well as a detailed overview of different types of evaluation planned. An overall description of the evaluation methodology is provided in section 3, which also describes the way feedback will be collected and communicated, in particular between users and technology partners. Technical testing is covered in Section 4, which will include all qualitative, automated and standardized tests that do not involve explicit user feedback. User-centered evaluation on the other hand is the focus of Section 5, outlining the plans to test the different levels of user testing in SELMA, including evaluation of the two use cases. Section 6 summarizes the planned evaluation in a timeline.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

The SELMA project is developing a StrEam Learning for Multilingual knowledge-trAnsfer (SELMA) platform.

The platform provides (1) a mechanism to seamlessly integrate different NLP (natural language processing) tools developed by multiple research partners in a variety of programming languages and environments; and (2) cloud-scalability of the platform to handle massive amounts of news data, well beyond what can be processed on a single or a few servers. It allows for the (3) continuous learning of deep learning models (4) model parameter sharing, and (5) fine-tuning of task models from user feedback. Two main use cases are targeted: Multilingual Media Monitoring and Multilingual News Production. In-depth evaluation is crucial to ensure a system that is reliable, stable, easy to use, flexible and expandable to make it sustainable and available to the media and research communities.

This document focuses on the evaluation strategy and planning for the SELMA project and describes targeted evaluation at different levels:

- individual components
- integrated platform and demonstrators
- targeted use cases and use case applications

We look at the objectives, the evaluation methodology, and include an overall description of the technical and user evaluation that is envisaged. We provide a detailed table of what evaluations are planned at the different levels and per partner, serving as a work sheet to track and plan evaluations throughout the project. A timeline puts the different phases into a timely context.

2. Evaluation Objectives and Plan

This section lists the main objectives of the evaluation activities which will be the targets to which performance will be directed and which will be related to the success criteria as set in D1.1 - Use Case Descriptions and Requirements.

2.1 Overall Evaluation Objectives

The objectives of all evaluation activities will strongly correlate with overall project objectives and will showcase if and how the project reached its set goals and objectives.

Overall, the goal is to develop a platform that is stable, easy to use, flexible and expandable. After the project, the tools and separate components are expected to be supported by consortium partners, to ensure sustainability, e.g. IMCS (Institute of Mathematics and Computer Science) and Priberam intend to maintain their demonstrators and make them available to other users. The tools can be further customized in terms of adding external engines or applications via API (application programming interface) or more language options, upon the client's request. Deutsche Welle (DW) envisages using services and components from the SELMA platform to enrich and improve its editorial production processes and its archiving and retrieval activities.

The project will start with proven prototypes that have previously been confirmed to work for media monitoring in the SUMMA project and the UX (User Experience) and production workflow developed for subtitling and voice-over generation in the plain X platform. The SELMA project partly builds upon these prototypes and extends them with continuous transfer learning capability from external data streams and user feedback, resulting in a system that becomes better with increased use, capable of ingesting massive amounts of different sources (news, internet feed, social media, etc.), and produce well-organized and topic-driven information that facilitates the propagation of key information to the end users. SELMA will also provide technology blocks for the related FTI (Fast Track Innovation) project Monitio (https://monitio-project.eu/).

The main objectives for the evaluation activities are the following:

#	Evaluation Objective
1	Evaluate the outcomes of the novel methods for training (and updating) machine learning/deep learning models for multiple speech and language tasks continuously.
2	Evaluate and benchmark the outcomes of the newly developed unsupervised multilingual language models for all 30+ project languages.
3	Evaluate the improvement of downstream tasks like entity recognition and linking, topic labelling, clustering, transcription, abstractive news summarization, automatic post-editing in all in all 30 languages.
4	Evaluate different clustering algorithms.
5	Evaluate outcomes of knowledge transfer across tasks in situations with asymmetrical amounts of resources between languages and tasks, particularly low resource languages.
6	Evaluate the newly developed data analytics methods and visualizations for improving the readability and access to information in order to boost and facilitate the decision-making process of media monitoring analysts and any global end-user in terms of accuracy and usefulness.
7	Evaluate functionality, usability and user acceptance for media monitoring workflow.
8	Evaluate functionality, usability and user acceptance of the multilingual content production workflow, particularly the multilingual transcription and translation models trained within the SELMA platform to enable an editorial production and content re-use workflow for 30 languages.
9	Evaluate overall media monitoring workflow for analytics for decision-making by media professionals
10	Monitor, validate and evaluate the outcome of the newly developed user feedback input and self-learning workflow for the improvement of the deep-learning model.
11	Evaluate whether the usage of the integrated workflows enabled by the SELMA platform will measurably improve the ease of multilingual content monitoring and creation. Evaluate the overall acceptance of the novel tools and workflows.

Table 1 *Evaluation Objectives*

2.2 Detailed Evaluation Objectives and Plan

In order to keep track of evaluations at the different levels and by the different consortium partners, a table in the form of an Excel sheet has been set up as a tracking tool. It lists the

components developed and evaluated within the project, with details on what aspects are the focus per partner and what kind of evaluation is planned.

Basic Component Overview

Component	**Partners involved in development and assessment**
1. Automated Speech Recognition (ASR)	LIA, FhG, IMCS, Priberam, DW/users
2. Machine Translation (MT)	LIA, FhG, IMCS, Priberam, DW/users
3. Abstractive summarization	Priberam, IMCS, DW/users
4. Named Entity Recognition (NER), Named Entity Linking (NEL), discovery	LIA, Priberam, ICMS, DW/users
5. Automatic post-editing of transcriptions and translations	LIA, FhG, Priberam, IMCS, DW/users
6. News clustering	Priberam, FhG, IMCS, DW/users
7. Topic detection	Priberam, FhG, IMCS, DW/users
8. Speech synthesis	LIA, FhG, Priberam, IMCS, DW/users
9. Alert system	Priberam, IMCS, DW/users
10. Indexation	Priberam, IMCS, DW/users
11. Search and visualization in User Interfacing	Priberam, IMCS, DW/users

12. Story segmentation	FhG, Priberam, IMCS, DW/users

Table 2 *Basic Component Overview*

This will be supplemented by a comparative evaluation of approaches such as performance of components with and without transfer learning to assess the impact of applying such novel approaches. These are detailed in the respective technical deliverables, e.g. D2.1 - Initial Progress Report on Continuous Massive Stream Learning and D3.1 - Initial Progress Report on Speech and Natural Language Processing.

The platform and components developed within the project – and their subcomponents and technologies – will be evaluated by technical partners and user partners for their accuracy, innovation and usefulness, and eventually measured against the KPIs (Key Performance Indicators) as described in D1.1 - Use Case Description and Requirements.

Key Performance Indicators

The table below lists the KPIs as they appear in D1.1:

Technology	**KPI**
Platform - Ingestion Scalability	Able to scale up to processing around 10M news articles/segmented video transcripts per day, given the computational resources.
Platform – User Scalability	The system should handle 500 simultaneous users for an installation of the platform.
Platform – Language Coverage	Processing models for Albanian, Amharic, Arabic, Bengali, Bosnian, Bulgarian, Chinese, Croatian, Dari, English, French, German, Greek, Hausa, Hindi, Indonesian, Kiswahili, Latvian, Macedonian, Pashto, Persian, Polish, Portuguese for Africa, Portuguese for Brazil, Romanian, Russian,

	Serbian, Spanish, Turkish, Ukrainian, Urdu. We will also include the two newly added DW languages: Tamil and Hungarian.
Component - Crosslingual Representations and Entity Linking	We will evaluate the representations on downstream tasks that make use of them, such as entity linking and text classification. We seek to outperform scores of the state of the art on standard offline multi-pass methods, improving over 86.6 micro F1 scores on end-to-end linking on the CoNLL YAGo dataset (Kolitsas et al., 2018).
Component - Summarization	We expect gains of at least 5% in ROUGE-1, ROUGE-2 and ROUGE-L F-scores over the state-of-the-art scores and improvements over 5% on factual correctness for abstractive summarization. A common dataset for evaluation is the CNN/DailyMail (Hermann et al., 2015) and New York Times dataset (NYT) (Sandhaus, 2008).
Component - Clustering	We expect improvements of at least 5% in the F1 metric.
Component - Segmentation	We aim at an improvement of 10% relative of the Diarization Error Rate (Bredin 2017) over state-of-the-art segmentation systems such as pyannote or the segmentation provided by Kaldi (Povey 2011), measured on the indomain data provided by Deutsche Welle.

Component - Speech Machine Translation	The system will be able to offer 2 new language pairs thanks to the use of D1.1 Use Case Description and Requirements 67 state-of-the-art end-2-end spoken machine translation approaches, especially for low resource speech translation.
Component - Automatic Transcription	The SELMA ASR system will outperform state-of-the-art results on some high resource language benchmarks. For instance, for French, on the ETAPE benchmark dataset, a reduction of at least 5% of word error rate is expected.

Table 3 *Technology and KPI*

Technology Readiness Level

Final assessment of the technologies will also look at the results ín terms of TRL (Technology Readiness Level). This will be based on the expected TRL improvement, as projected in the SELMA Description of Work.

Technology	**TRL (2020)**	**Expected TRL (2022)**
Punctuation Recovery	5	7
Speaker Diarization	6	8
Speaker Recognition	6	8
Rich Automatic Speech Recognition (including named entities)	6	8*

Text Machine Translation	7	9*
Expressive and Personalized Voice Synthesis	5	7
Speech Machine Translation	5	7
Topic Labeing (from crosslingual transfer)	4	6
Named Entity Recognition and Linking	6	8
Abstractive Summarization	3	6
Integration Platform (NLP Components and UX)	6	8
Integration Platform (Learning/Training of NLP and Automatic Redeployment)	3	7
*depending on the target languages and language pairs		

Table 4 *Technology Readiness Level Overview*

Detailed Component Evaluation Plan

A detailed component evaluation plan in the form of an extensive online Excel table is the central working tool for SELMA evaluation. It serves as the overall evaluation tracking and planning tool, ensuring an overview and proper coordination of evaluation activities for the entire consortium at technical as well as user level. It is a live document and will be continuously updated and expanded throughout the project.

Below is a screenshot of the Excel sheet containing the Evaluation Plan and Tracking sheet:

Component Name	Partners	Component Level Testing	Integrated Platform Level	Demonstrator Level
ASR (Automatic Speech Recognition)	Tech: LIA	WER (Word Error Rate) scoring, shared task evaluation		podcast, all apps
ASR enriched with speaker recognition		Evaluation metric for ASR with speaker recognition: WER + speaker tracking evaluation techniques, used in NIST-SRE evaluation (https://sre.nist.gov/)		
ASR without punctuation, ASR enriched with punctuation				
ASR	Tech: IMCS	WER scoring	functional, workflow, scalability tests	functional, workflow, scalability tests
ASR	Tech: Priberam		Integration testing on plain X and Insight	
ASR (German, English and more)	Tech: FhG	WER scoring, shared task evaluation - Evaluation metric for ASR: WER (Word Error Rate) - evaluation of automatic post-editing		
ASR for low-resourced languages		Using transfer learning techniques to optimize ASR for low-resourced languages. Metrics: WER scoring, shared task evaluation, evaluation of automatic post-editing		
ASR	User: DW	Upon technology partner request		Insight, plain X, podcast - quality assessment: user rating, benchmarking
MT (Machine Translation)	Tech: LIA			
Text-to-text translation		Metrics for text-to-text translation: BLEU		
Speech-to-text translation		Metrics for speech-to-text translation: BLEU		
Speech-to-speech translation		Metrics for speech-to-speech translation: BLEU for the transcription of the target language, intelligibility, perceptual evaluation (there are still open research issues for techniques of evaluation of speech- to-speech translation)		
MT	Tech: IMCS		functional, workflow, scalability tests	
MT	Tech: Priberam		Integration testing on plain X and Insight	
MT	User	Overall MT output quality assessment, gap filling, direct assessment Compare speech-to-speech output to traditional method (ASR + MT) and see if prosody has improved.		User rating: accuracy and usability - user rating, benchmarking MT: Quality assessment in plain X, Insight and podcast Speech-to-speech: in podcast and plain X. Compare quality output to traditional method (ASR + MT). See if prosody has improved.
Abstractive news summarization	Tech: Priberam	ROUGE and new methods to be explored, using a standard dataset. Human evaluation on top of this is essential to avoid hallucination and to improve factuality.	Integration testing on plain X and Insight	
Abstractive news summarization	Tech: IMCS		functional, workflow, scalability tests	
Abstractive news summarization	User	Upon technology partner request		Insight: Assess level of factuality and hallucination by means of annotation of answering questions based on the summaries. This uses a small subset of the larger dataset and questions are manually created for the subtestset. Evaluation will be through a customized UI.
Named entity recognition (NER) and linking (NEL) and discovery	Tech: LIA	NER and NEL component for ASR - Evaluation metrics: CER (concept error rate), CVER (concept value error rate)		
Named entity recognition (NER) and linking (NEL) and discovery	Tech: Priberam	Priberam NER and NEL component for text: Metrics: F1. Goal is to improve NER and NEL using feedback from the datastream. Comparison of output from the two methods: basic method without datastream and new method with datastream. Look for improvements without retraining the models. Assess user input/corrections logs on NER/NEL. The system should learn to disambiguate once a user has given a correction.		
Named entity recognition (NER) and linking (NEL) and discovery	Tech: IMCS		Functional, workflow, scalability tests	
Named entity recognition (NER) and linking (NEL) and discovery	Tech: Priberam		Integration testing on plain X and Insight	
Named entity recognition (NER) and linking (NEL) and discovery	User	Upon technology partner request		Insight User evaluation = user feedback (logs) and user rating Accuracy assessment on Insight: checking if the names are correct and useful from user side. Through user feedback by user corrections of NER/NEL (in Insight already in place). Measures percentage of corrections needed. Goal: later on the system should learn to disambiguate once a user has given a correction.
				User rating of accuracy and usefulness.
Automatic post-editing of transcriptions and translations	Tech: LIA, Priberam	Named entities post-editing (correcting named entities in text output of ASR and MT).	Integrated testing on plain X and Insight	
		Generic automated post-editing (correcting the text output of ASR and MT, for instance applied to another domain)		
		Measure the quality performance without and with post-editing.		
Automatic post-editing of transcriptions and translations	Tech: IMCS		Functional, workflow, scalability tests	
Automatic post-editing of transcriptions and translations	Tech: FhG	With focus on entities, in particular transfer from high-resource languages to low-resource languages (for entity recognition. Metrics: WER		
Automatic post-editing of transcriptions and translations	User	At technology partner request		Insight and plain X Quality assessment: checking accuracy by means of editorial post-editing of named entities and text output of ASR and MT. User logs and user rating
News Clustering	Tech: Priberam	F1 and B-cubed F1 scores, using a standard dataset	Integration testing on Insight	
News Clustering	Tech: IMCS		Functional, workflow, scalability tests	
News Clustering	Tech: FhG	News classification and clustering. F1 scores, starting from standard dataset and compiling customized dataset		
News Clustering	User	At technology partner request		Insight User rating: accuracy and usability User evaluates whether the stories are well separated and if it is useful. Does the clustering simplify the task of monitoring a stream of documents?

				Multilingual clustering: documents coming from different languages will be clustered
Topic detection	Tech: Priberam	Insight: evaluation F1 using several datasets, including Priberam's	Integration testing on Insight	
Topic detection	Tech: IMCS		Functional, workflow, scalability tests	
Topic detection	Tech: FhG	Using the same data as news classification and putting topic labels onto it. Metrics: F1 scores, starting from standard dataset and compiling customized dataset		
Topic detection	User	At technology partner request		Insight User rating: accuracy and usability Multilingual topic detection: evaluate whether the detected topics are right or wrong Rating (e.g. percentage of correct topics: number of correct topics out of total).
Speech synthesis	Tech: LIA	Evaluation metric: qualitative evaluation, user evaluation (there are still open research issues for techniques of evaluation of speech synthesis)		
Speech synthesis	Tech: IMCS		Functional, workflow, scalability tests	
Speech synthesis	Tech: Priberam		Integration testing on plain X	
Speech synthesis	Tech: FhG	Support TTS with a good pronunciation - collaborating input for the pronunciation dictionary and experiment with TTS from another partner). Qualitative evaluation (before and after correction) and user testing.		
Speech synthesis	User	At technology partner request		podcast, plain X User rating: accuracy, fluency and usability
Speech synthesis - Expressivity	Tech: LIA	Evaluation metric: New evaluation metrics based on clustering, perceptual evaluation		
Speech synthesis - Expressivity	User	At technology partner request - User rating: expressivity		podcast, plain X User rating: expressivity, flexibility, and usability Evaluate if the synthetic voice output approximates the expressivity of the original voice.
Alert System (Breaking News Alerts)	Tech: Priberam		Functional testing	
Alert System (Breaking News Alerts)	Tech: IMCS		Integration testing	
Alert System (Breaking News Alerts)	User			Insight User rating Functional user evaluation: Do I get breaking news alerts as expected? Usability testing: Are the breaking news alerts useful?
Indexation	Tech: Priberam	Indexing functional quality testing, e.g. accuracy	Indexing quality testing at system level, e.g. about speed of indexation, measuring how the platform behaves	
Indexation	Tech: IMCS		Integration testing	
Indexation	User		quality assessment	Functional user evaluation: Does the system provide the user with indexed data? Usability testing: quality assessment to see if users can find things efficiently
Search and Visualization in User Interfacing	Tech: Priberam	UI testing - functional testing	Integration testing on Insight	
Search and Visualization in User Interfacing	Tech: IMCS		Functional, workflow, scalability tests	
Search and Visualization in User Interfacing	User			Insight Quality assessment: user rating Functional and Usability testing Can users (easily) find targeted data to perform their editorial tasks. Is is useful in their work?
Story Segmentation	Tech: FhG	Speaker Diarization - Metrics: DER (Diarization Error Rate) - for speaker recognition: accuracy measurement		
Story Segmentation	Tech: IMCS		Functional, workflow, scalability tests	
Story Segmentation	Tech: Priberam	Focus on topical analysis	Integration testing on Insight	
Story Segmentation	User	At technology partner request		Insight Quality assessment: user rating: are stories properly segmented from a reader point of view?
Full workflow	Tech: IMCS		Functional, workflow, scalability tests	
Full workflow	Tech: Priberam		Integration testing on Insight and plain X Both platforms have Google Analytics to track usage	
Full workflow	Tech: FhG			Assessment of how the FhG models perform in the demonstrators (functional testing)
Full workflow	User			Insight and plain X User evaluation of the platform as a whole with the entire workflow as described in D1.1 of WP1 : all aspects of plain X : (ASR - MT - subtitling - voice synthesis) + all aspects of Insight. Feedback will be provided using tickets on Trello (Feedback from the users will be filtered, clustered and compiled in Trello as bugs or other feedback)

Figure 1 *Screenshot of Excel Sheet with Detailed Evaluation Plan*

3. Evaluation Methodology

This section presents an overview of the underlying methodology to the evaluation activities and evaluation planning.

3.1 Overall Evaluation Methodology

The overall evaluation methodology will have a dual, iterative, modular and technology-specific approach. Dual, in terms of involving user-centered as well as technical evaluation, iterative as each evaluation phase and individual activity will provide direct insights and feedback for development and improvement, modular and technology-based as each individual functionality and technology advancement could potentially be assessed and evaluated individually as well as part of the overall workflow or integrated platform.

Several components developed within the SELMA project will also be integrated and tested in other projects and integrated platforms such as Monitio (https://monitio-project.eu/, an EU FTI project in which Priberam and DW are participating and which uses the Insight platform), plain X (integrated editorial production platform developed by Priberam and Deutsche Welle for semi-automated transcription, translation, subtitling and voice-over, being rolled out in Deutsche Welle) and others. A strong emphasis will be put on technical evaluation and validation of the newly developed models and benchmarking for industry standards to see if the project reached its objectives and KPIs.

Technical tests will be conducted by the technical partners, on components, technologies, and specific functionalities before and during the respective user tests to provide consistent functionality.

The technical tests will primarily be reported in the respective technical deliverables. The current document will also provide a general overview of the technical evaluation activities planned throughout the project as they play a crucial role in the overall evaluation framework. Such an overview is needed to ensure technical and user testing are covered for all components and technologies envisaged and to have a working tracking tool. User-centered evaluation will put the role of the end users and their workflows and needs into focus. To provide the best possible outcome for this purpose, prototypes, modules and functionalities

will have to be tested in controlled near-real-life scenarios, in which the test users will be given tasks to perform. Platform testing will provide an overview of real-life stability, integration and functionality of the project's outcome.

User Acceptance, which includes essential aspects such as usability, accessibility, perceived usefulness, technology uptake barriers and benefits, will be a major part of user-centered evaluation activities. The different aspects of SELMA output will be tested and evaluated using several different user-centric evaluation tools and methods, including but not limited to surveys, interviews, questionnaires, focus groups and platform usage patterns.

3.2 Test User Recruitment and Onboarding

The test user group will primarily consist of DW media professionals. Internal test users will be recruited based on specific activities, test phases and languages. There will, however, also be an internal core test user group which will be engaged throughout the project and be able to evaluate the overall progress of the project results. The internal core test user group will consist of DW Innovation department staff as well as selected DW editors and other media professionals and NLP specialists.

Test user recruiting will have the goal to correspond with the envisioned project personas, taking into account diversity in terms of roles and experience with NLP-based editing and content creation tools. The consortium will actively seek to establish a gender-balanced test user group.

The project has also set up a User Group, with a smaller subset serving as an Advisory Board. They have a consulting function for the Consortium.

Members of the User Group include professionals from major broadcasting organizations such as EBU, SWR, ARTE, RAI, BBC, EuroNews, Prisa, and some research organizations such as the University of Edinburgh and the University of Tilburg. Current users using the Insight prototype now being exploited by Priberam, e.g. EMBRAER, AICEP and LANDAU Media, will also take an active part in the validation of SELMA. The test user group may gradually expand with expanded dissemination efforts and attempts to mirror the geographic diversity of the consortium.

The members of the User and Advisory Board will be invited to project-organized workshops, in which they will review the project's activities and results, identify the strong and weak points with respect to the objectives of the project (with emphasis on the innovation objectives), and provide recommendations. Furthermore, the members of the User and Advisory Board help us maximize our industry outreach, serving as links between the consortium and external key industry players. The recruiting of potential candidates will continue as the project progresses.

User Days will be set up as main vehicles for the onboarding and engaging of interested external parties, to show the results and create a platform for all interested parties to try out the system and become a beta tester and provide feedback. We will also set up activities for possible external contributors who may work on add-ons to the platform or experiment with the project outcome especially for the Open-Source aspects of the project. User Days in the later project phases will focus on exploitation and encourage cooperation beyond the project's timeframe.

3.3 Feedback Provision and Communication

In order to ensure a proper – and continuous – feedback flow between users and developers and between the different technical partners, the following tools are foreseen to provide and track feedback:

From users to developers and integrators:

- User Project Team evaluation
    - To test demonstrators for functionality testing, user interfacing and usability: feedback button on UI + Trello tickets
    - For component testing: customized UI for feedback + Trello tickets
- Editorial user evaluation
    - To evaluate demonstrators for functionality testing, user interfacing, usability and user acceptance testing: feedback button on UI + feedback to User Project Team - this will be done through customized forms, surveys, interviews, for instance. The user feedback will be subsequently filtered and submitted as Trello tickets.

    - Feedback from user post-editing transcriptions, translations, summaries, named entities, etc.: feedback provided and collected through demonstrator use.
    - Component testing can be assigned to selected editorial users: customized UI or feedback forms.
- Automated feedback tracking: user logs as implemented and analyzed by the platform developers.

Between technology partners:

- Feedback on integrated platform and demonstrator level testing: Trello tickets
- Feedback on specific technologies and components: Trello or customized feedback systems

4. Technical Testing

This section describes technical testing planned for individual components as well as integrated platforms and demonstrators. In the evaluation work package and deliverables, we keep an overview of the different types of assessment, overall results, and relations between and impact of evaluation activities. Specific technical details are reported in the respective technical deliverables, e.g. D2.1 - Initial progress report on continuous massive stream learning and D3.1 - Initial progress report on speech and natural language processing.

4.1 Unit Testing

SELMA NLP components developed by the University of Avignon (LIA), Fraunhofer (FhG), Priberam, and IMCS will initially be tested on their own, sometimes with a special UI. The purpose is to validate that the software of the component performs as expected. This first-level testing is done by the developing partner and precedes integration testing.

The 12 components listed in Section 2.2, Table 2 - Basic Component Overview, are envisaged to be developed and evaluated:

Component-level testing has been started and is reported in detail in the respective technical deliverables. For instance, initial findings of progress using different approaches for crosslingual clustering and news classification, NER and NEL can be found in section 4 of D2.1 - Initial progress report on continuous massive stream learning. Initial results as to ASR, NER in ASR, speech translation (AST), spoken language understanding (SLU), speech synthesis are reported in sections 2, 3, 4 and 5 of D3.1 - Initial progress report on speech and natural language processing.

4.2 Integration Testing

Integration testing will be at the overall integrated platform level, as well as specific demonstrator level by the relevant technical partners. They will serve as a second-level evaluation of the components' performance. This testing covers for instance integration stability and performance assessment, scalability and stress tests.

The integrated software testing envisaged in SELMA includes:

1. Use Case 0, the light-version demonstrator developed and tested by IMCS, introduced in D4.1 - Platform architecture and API documentation, to simplify testing of NLP components before they are integrated into UC1 and UC2.
2. Use Case 1 , with the Insight and Monitio demonstrators, developed and tested by Priberam.
3. Use Case 2 , with the plain X demonstrator, developed by Priberam and DW, tested by Priberam.
4. Overall integration by IMCS, as part of WP4, the "glue" between all components, dealing primarily with UC0 and the scalability of NLP components to 10M docs/day, tested by IMCS.

Integration testing will be reported in detail in the prototype deliverables of years 2-3 (D1.2 - Initial Prototype Report, D1.3 - Intermediate Prototype Report, D1.4 - Final Prototype Report); the integrated platform deliverables (D4.2 - Initial platform release with the primary NLP pipeline, D4.3 - Intermediate platform with continuous massive stream learning NLP capabilities, D4.4 - Final platform release with full continuous massive stream learning capabilities); and the demonstrator-focused deliverables (D4.5 - Demonstrator for use case one, D4.6 - Demonstrator for use case two).

5. User Evaluation

This section provides an overview of the approach in terms of user evaluation. As described in Section 3, qualitative evaluation activities will be carried out on multiple levels. There will be evaluation activities at the component level and we look at further use of the components for other projects or platforms including plain X and Insight. At the demonstrator level, we have evaluation activities at the scenario, use case and use case application level.

The user-centred evaluation for the two main SELMA use cases (media monitoring and media news production) will include the testing of all features that should be available for the users as established in the initial requirements and associated usage examples and User Scenarios (see D1.1 - Use Case Description and Requirements).

Moreover, the evaluation activities will assess usability and question whether the test users see a general benefit in using the platform for their day-to-day work. The user evaluations for the use cases will be performed in individual pilot tests based around a concrete assignment.

User observation, questionnaires, and interviews will be used to collect measurable results in terms of usability. We will work towards five characteristics of usability, the five E's: effective, efficient, engaging, error-tolerant and easy to learn - and adhere to the description of these five characterics as provided in https://www.wqusability.com/:

- **Effective:** Effectiveness is the completeness and accuracy with which users achieve specified goals. It is determined by looking at whether the user's goals were met successfully and whether all work is correct.
- **Efficient**: Efficiency can be described as the speed (with accuracy) in which users can complete the tasks for which they use the product.
- **Engaging**: Ensure the platform is pleasant and satisfying to use.
- **Error-tolerant**: The ultimate goal is a system which has no errors, but most systems are far from perfect. An error tolerant program is designed to prevent errors caused by the user's interaction, and to help the user in recovering from any errors that do occur.

- **Easy to learn:** It allows users to build on their knowledge without deliberate effort. This includes general helpfulness, access to instructions and predictability, i.e., place information or controls where the user expects it to be.

5.1 User Evaluation at Component Level

On the component level, DW will perform specific user testing upon request by the technology partners. For some components, technology partners will provide a customized UI, as is the case for NER and NEL assessment. User tests will comprise an overall quality and usefulness assessment through rating and specific test formats such as gap-filling exercises and direct assessment. The latter will be selectively applied, for instance for machine translation or summaries, involving native speakers. This fits within the Deutsche Welle NLP benchmarking effort, as described in Section 5.5. Similar activities will be done for the other individual NLP tasks, including speech-to-speech translation or named entity recognition and linking.

5.2 User Evaluation for Use Case 1 - Media Monitoring

User Evaluation will focus on the use cases, and the subordinate use case applications as described in section 3 of D1.1 - Use Case Description and Requirements. These will be tested primarily in the demonstrator environment made available to the users.

The first use case, media monitoring, focuses on extensive, multilingual media monitoring by daily observation of web feed content, like RSS feeds and website site maps which give access to written articles, video and audio content in multiple languages. The user can choose from available feeds and assigns them to a group, save the latter for easy reference as well as aggregate the media items contained in these groups. An automatic analysis identifies clusters of media items that are related by topic or another common attribute. The goal is to assist media professionals in day-to-day information gathering spanning multiple languages and media outlets. Each individual step in the foreseen workflow will be evaluated with media professionals in realistic settings in terms of functionality, accuracy, usefulness and acceptance for media monitoring and decision making.

One use case application evaluation of news monitoring is to assess the ability to analyze an arbitrary group of articles with respect to the diversity of their content. In order to achieve this, recognized named entities will be augmented with the addition of metadata from Wikidata. More concretely, the following fields are added where available: sex and gender, country of citizenship, ethnic group, sexual orientation, medical condition, religion, educated at, date of birth. The addition of these fields will make it possible to derive diversity-specific statistics on article groups. The evaluation will focus on the accuracy of the calculated statistics, as well as on their usefulness as part of the Insight Platform demonstrator.

5.3 User Evaluation for Use Case 2 - News Production

The news production use case focuses on facilitating content creation.

One such use case application foresees a system which observes (DW or other broadcast) content output in a large variety of languages as well as the output of a number of configurable external news outlets and creates newsletters for company-internal consumption, providing regular updates on basic or advanced content analysis and producing summaries.

Another major content creation application focuses on the ability to transcribe, translate and subtitle media items (text, video and audio), and provide voice-over using synthetic speech. SELMA aims to support all of Deutsche Welle's 32 languages. The user evaluation focuses on the quality of the produced transcriptions and translations. This is currently being evaluated in the plain X demonstrator.

Another example is the audio podcast creation use case application. The system proposes a list of storylines that were trending based on a set of configurable search parameters (countries, topics, named entities), over a configurable period. The journalist is able to select suitable storylines from those, and the system assists in transforming them into audio news bulletins, using summarization and voice synthesis techniques. Each individual step in the foreseen workflow will be evaluated with media professionals in realistic settings in terms of functionality, accuracy, usefulness and acceptance for content production.

This use case application is currently under development. Two different approaches are targeted.

a) the assisted creation of a news podcast in Brazilian Portuguese

b) the translation of an existing podcast into another audio language

In the case of a), a demonstrator will be built with the purpose to serve as a user interface to test the functionality of the relevant technical components. It will use the platform to group relevant articles, create storylines and summarize them. The journalist will be able to edit the text before instructing the system to synthesize a news podcast based on the edited summaries. The evaluation will concentrate on assessing the acceptance of such a synthetic podcast with the end user, as well as calculating the increase in productivity in comparison to a conventionally produced version of a news podcast in the same language.

In the case of b), a second demonstrator will be built that assists the journalist in selecting a suitable DW podcast (most probably in English) and convert it into another language. Again, the acceptance with the end user of such a podcast will be evaluated, together with the calculation of the gain of productivity on the editorial side.

5.4 Evaluation at User Scenario Level

Functionality testing at scenario level is a crucial part of user testing. This will cover the 22 scenarios as foreseen in D1.1 - Use Case Description and Requirements. An overview of the scenarios is depicted in the graph below (Figure 2 – Overview User Scenarios), with details of each scenario in Table 3.

The scenarios are functional areas identified as being relevant to SELMA and based on the personae and workflow descriptions as defined during the requirements process. The Scenario Model (Figure 2) defines these scenarios and their interaction with both SELMA and the relevant personae, whilst the individual scenario descriptions (Table 3) define the functional path within each of these scenarios.

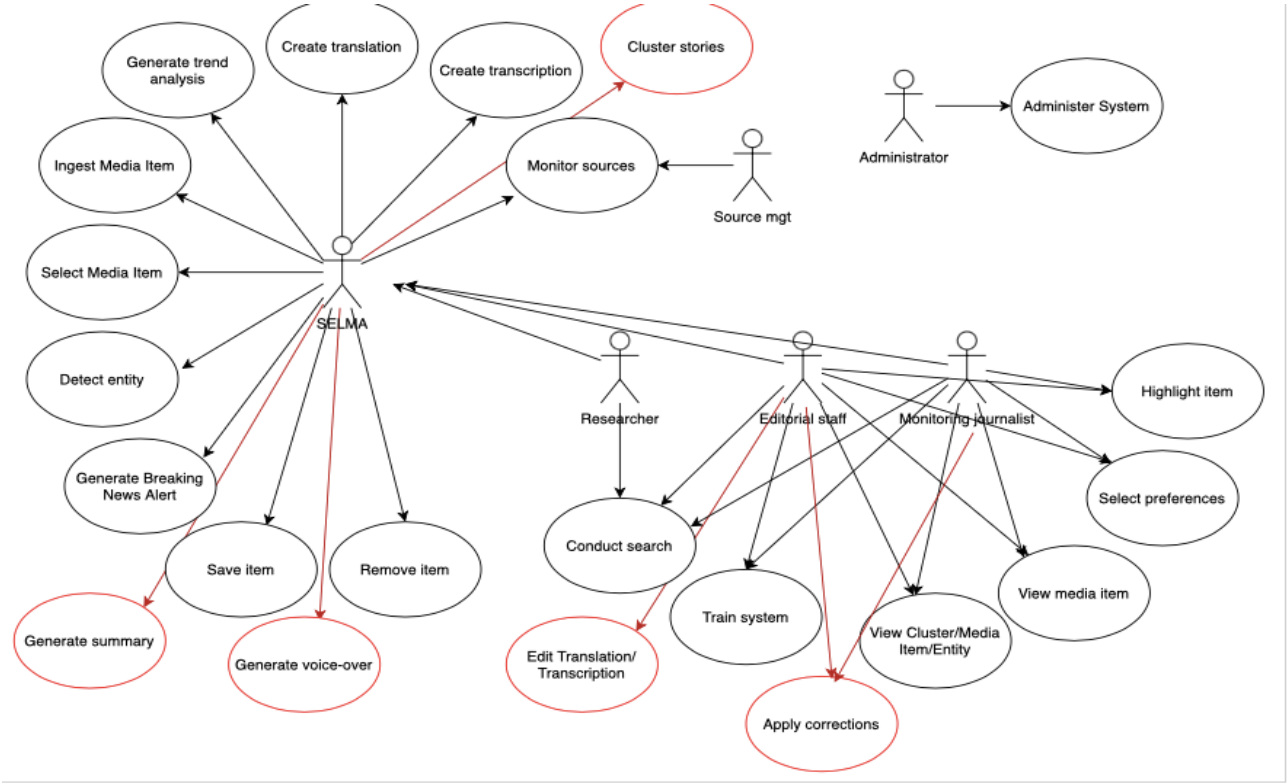

Figure 2 *Scenario Model Showing an Overview of User Scenarios*

At scenario level, all individual functionalities will be evaluated in multiple rounds to determine and measure their usability, accuracy and improvement over time. The scenarios can be seen as the smallest functionality unit and hence evaluated individually or as part of a use-case-related workflow and integrated in multiple demonstrators and user interfaces. By breaking evaluation activities down to the scenario level, we will be able to compare the outcome of different types of testing and evaluation activities.

Table 3 (User Scenarios) below lists the targeted scenarios, provides a brief description for each and the focus of the evaluation.

User Scenarios in Detail

#	Scenario ID	Scenario Description	Focus Evaluation
1	Monitor Sources SEL-Sc-001	The user and language team specifies the input source(s) they wish to monitor through the system.	Functionality

2	Ingest Media Item SEL-Sc-002	The system ingests media items from the sources.	Functionality
3	Select Media Item SEL-Sc-003	The system selects media items and shows them to the user based on specific preferences.	Functionality
4	Detect and Link Entity SEL-Sc-004	The system detects an entity and links it to other media items or clusters from the sources being monitored based on preferences specified by the user.	Functionality, Accuracy, Gradual Improvement
5	Generate Breaking News Alert SEL-Sc-005	The system generates breaking news alerts based on the preferences set by the user.	Functionality, Relevance
6	Create Transcription SEL-Sc-006	The system creates a transcription for an individual AV media Item.	Functionality, Accuracy, Gradual Improvement
7	Create Translation SEL-Sc-007	The system creates a translation for an individual AV media item.	Functionality, Accuracy, Gradual Improvement
8	View Cluster/ Entity SEL-Sc-008	The user views the details of a cluster and/or an entity.	Functionality
9	View Individual Media Item SEL-Sc-009	The user views an individual media item in relation to a cluster or entity.	Functionality
10	Select Preferences SEL-Sc-0010	The user sets their preferences in the system.	Functionality, Usefulness
11	Conduct Search SEL-Sc-0011	The user can search for an item in the system.	Functionality, Relevance of results
12	Save Cluster / Individual Media	The user can save a cluster or an individual media item in the system where it is stored for more than a	Functionality

	Item SEL-Sc-0012	predefined set of time.	
13	Remove Item SEL-Sc-0013	The user can remove an individual item and/or a cluster (with all its associated media items) from their view.	Functionality
14	Train System SEL-Sc-0014	The user can train the system in relation to the cluster generation.	Functionality, Accuracy, Gradual Improvement
15	Highlight Item SEL-Sc-0015	The user can highlight an item to make it visible to other members of the user's team.	Functionality
16	Generate Trend Analysis SEL-Sc-0016	The system carries out a trend analysis and presents the results to the user.	Functionality, Accuracy, Usefulness, Gradual Improvement
17	Administer System SEL-Sc-0017	The System Administrator carries out various activities to administer the system.	Functionality
18	Group Media Items into Clusters SEL-Sc-0018	The system clusters media items based on the preferences set by the user.	Functionality, Relevance and Accuracy, Usefulness, Gradual Improvement
19	Generate Summary SEL-Sc-0019	The system generates a summary for each media item.	Functionality, Relevance and Accuracy, Usefulness, Gradual Improvement
20	Generate Voice-Over SEL-Sc-0020	The system generates a voice-over for a transcription and/or translation of a media item on demand.	Functionality, Accuracy and Expressiveness, Gradual Improvement
21	Edit Transcription/Translation	The user can edit and correct the transcription and the translation. It is possible for 2 users to edit a transcription/translation simultaneously.	Functionality, Ease of use

	SEL-Sc-0021		
22	Apply Corrections SEL-Sc-0022	The system applies the corrections made by the user to the rest of the single media item or its cluster as defined by the user.	Functionality, Accuracy, Gradual Improvement

Table 5 *User Scenarios in Detail*

5.5 DW NLP Benchmarking

DW will perform complimentary benchmarking for ASR (automated speech recognition) and MT (machine translation) through an in-house process that uses established benchmarking tools for NLP (natural language processing) engines, including the ASR benchmarking tool developed by the EBU (European Broadcasting Union), and this for all 32 Deutsche Welle languages. Initially, 10 priority languages have been selected (with a mix of low-resource and high-resource languages): Arabic, Chinese (both versions), Greek, Hindi, Indonesian, Pashto, Persian, Polish, and Turkish.

The rating of each engine will consist of an automated evaluation, supplemented by a human evaluation. The main automated metrics that will be used to assess the quality of each engine will be the WER (Word Error Rate) for ASR and the BLEU (BiLingual Evaluation Understudy) score for MT (machine translation). Engines used in the integrated platform, whether third-party engines, such as Google or Azure, or consortium partner components, will be included in the evaluation for comparison purposes.

The human evaluation will be collected through questionnaires and will focus on the rating of other parameters, such as the use of idioms, punctuation, entity spelling accuracy, etc. The goal is to provide a consistent and comparable result across languages. This will be achieved by using one video to be translated in all 32 DW languages. For selected high-resourced languages, benchmarking will be extended using several videos with different settings (speaker's accent, variety of topics, background noise, etc).

6. Timeline

Below we present an estimated broad timeline, summarizing our planned evaluation and providing a broad overview. This timeline will be updated as needed.

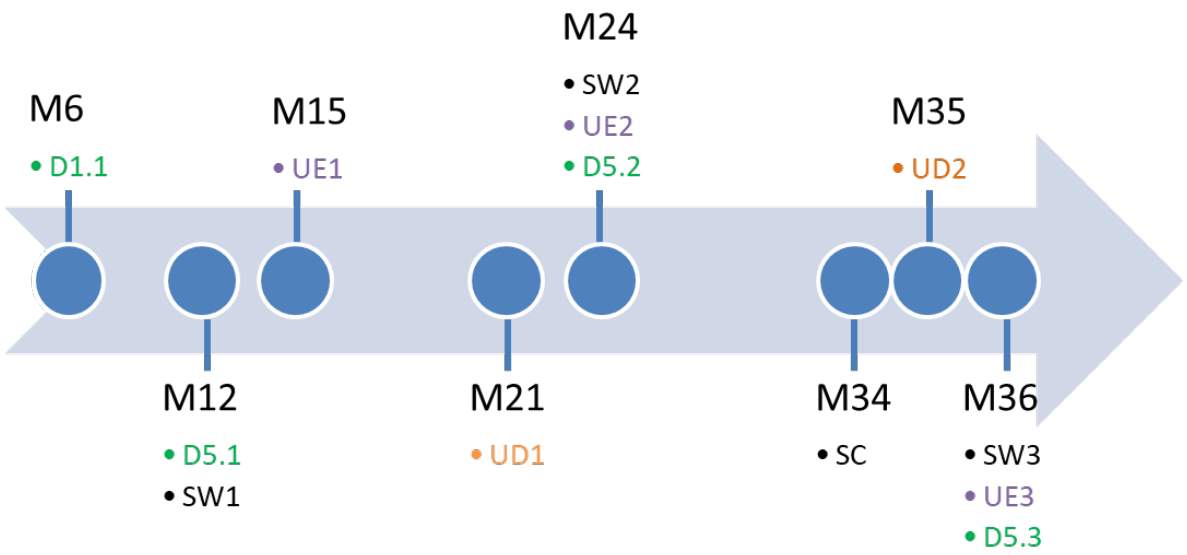

Figure 3 *Broad Timeline of Evaluation Activities Planned*

Timeline Legend:

- D = deliverable
- SC = scalability testing
- SW = software release
- UE = user evaluation (usability)
- UD = user day

The timeline above shows the milestones and deliverables according to the project plan. The actual evaluation will depend on the availability of the software and the need for specific

testing. Although user evaluation is marked at particular stages (M15, M24 and M36), it will take place throughout the project. For instance, component testing is done in coordination with or at the request of technology partners. Functionality testing on the platform will also be done as soon as certain features are implemented and available for testing, allowing evaluation over a longer period of time and providing regular and timely feedback in a consistent manner; at DW this will initially be performed by the DW Innovation Team, also use case applications will be assessed as available. Official, formal usability testing will be done during specific intervals with editorial staff at DW and external users from DW and Priberam's network, inside or outside of the User and Advisory Board. The last stage, focusing on user acceptance of the tools, should primarily take place between M31 and M35 to allow a feedback flow to the developers.

The User Days will serve as a main channel of outside involvement and include the open-source community and allows for user feedback collection.

7. Conclusion

This document presents an overview of the evaluation activities planned within the SELMA project. It has been compiled with contributions from all consortium partners and relates to other deliverables, including D1.1 - Use Case Descriptions and Requirements, D2.1 - Initial progress report on continuous massive stream learning, and D3.1 - Initial progress report on speech and natural language processing.

The overall methodology as to evaluation and the different types and levels at which evaluation is foreseen is outlined.

A detailed evaluation overview sheet serves as central evaluation planning and tracking tool, to ensure that all participants in SELMA are aware of planned, ongoing and completed testing, at component as well as integrated platform and demonstrator level, thus facilitating collaboration between technology partners and users alike.

Details are provided as to the process and procedures for technical testing on the one hand and user evaluation on the other. Also, various envisaged forms of feedback resulting from assessment are covered.

The input of this report (D5.1 - Evaluation Plan) will guide all partners in the evaluation process. This deliverable will be followed by two other reports in this series: D5.2 - Interim Evaluation Report and D5.3 - Final Evaluation Report.

8. Annex

8.1 Acronyms

Below is a list of acronyms that are used in this deliverable.

Acronym	Expansion
API	Application Programming Interface
ASR	Automated Speech Recognition
BBC	British Broadcasting Corporation
BLEU	BiLingual Evaluation Understudy (measurement for MT)
Dx	Deliverable x
DW	Deutsche Welle
EBU	European Broadcasting Union
FhG	Fraunhofer Gesellschaft
FTI	Fast Track Innovation
IMCS	Institute of Mathematics and Computer Science
KPI	Key Performance Indicator
LIA	Laboratoire Informatique d'Avignon
Mx	Month x
MSx	Milestone x
MT	Machine Translation
NEL	Named Entity Linking
NER	Named Entity Recognition

NLP	Natural Language Processing
NYT	New York Times
PRIB	Priberam
RAI	Radiotelevisione Italiana
RIA	Research and Innovation Action
SC	Scalability Testing
Sc	Scenario
SEL	SELMA
SELMA	Stream Learning for Multilingual Knowledge Transfer
SW	Software Release
SWR	Südwestrundfunk (German broadcaster)
ToC	Table of Contents
UC0	Use Case 0
UD	User Day
UE	User Evaluation
UI	User Interface
UX	User Experience
WER	Word Error Rate (measurement for ASR)