Research and Innovation Action (RIA) H2020 – 957017

SELMA

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D5.2 Interim Evaluation Report

Work Package	5
Responsible Partner	Deutsche Welle
Author(s)	Peggy van der Kreeft
Contributors	Andreas Giefer, all partners
Reviewer	Guntis Barzdins
Version	V1.0
Contractual Date	31 December 2022
Delivery Date	22 December 2022
Dissemination Level	Public

Version History

Version	Date	Description
0.1	03/11/2022	Initial Table of Contents (ToC)
0.2	1/12/2022	Initial Input
0.3	8/12/2922	Input from Technical Partners
0.4	18/12/2022	Internal Review
1.0	22/12/2022	Finalization and Submission

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 update of the evaluation efforts within the SELMA project, according to the previously established Evaluation Plan. It describes what has been done in terms of assessing the targeted platforms and use cases, as well as individual components. It showcases what technology partners and user partners have done and are collaborating in this respect.

This report describes the overall progress on evaluation in section 2. Section 3 refers to the work done on technical testing by the technology partners, on the 12 SELMA components and the three integrated platforms (SELMA Open-Source Platform, Monitio and plain X).

Table of Figures

Table of Tables

1. Introduction

The SELMA project is developing a StrEam Learning for Multilingual knowledge-trAnsfer (SELMA) platform, integrating different NLP (natural language processing) tools. More details on the objectives were given in D5.1 - Evaluation Plan.

This document focuses on the execution of the evaluation plan for the SELMA project, as described in D5.1 - Evaluation Plan. We will show progress in the evaluation in Y2 done at different levels:

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

Following the strategy outlined previously in the Evaluation Plan, we provide the updated sheet that shows progress on the different components and platforms.

2. Evaluation Plan and Progress Made

This section provides a broad overview of the evaluation activities in Year 2.

The principal objective is to ultimately develop a platform that is stable, easy to use, flexible and expandable.

As outlined in the Evaluation Plan, SELMA partly builds upon proven prototypes, in particular the SUMMA platform for monitoring and the plain X platform for content creation and adaptation. It extends these 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.

The main evaluation objectives we are pursuing are listed in the following table, with an indication of those objectives that have been evaluated during the reporting period.

#	Evaluation Objective	In process
1	Evaluate the outcomes of the novel methods for training (and updating) machine learning/deep learning models for multiple speech and language tasks continuously.	Y
2	Evaluate and benchmark the outcomes of the newly developed unsupervised multilingual language models for all 30+ project languages.	Y
3	Evaluate the improvement of downstream tasks like entity recognition and linking, topic labelling, clustering, transcription, abstractive news summarization, automatic post-editing in all 30 languages.	Y
4	Evaluate different clustering algorithms.	Y
5	Evaluate outcomes of knowledge transfer across tasks in situations with asymmetrical amounts of resources between languages and tasks, particularly low resource languages.	Y
6	Evaluate the newly developed data analytics methods and visualizations for improving the readability and access to information in order to boost and	Y

	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.	Y
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.	Y
9	Evaluate overall media monitoring workflow for analytics for decision-making by media professionals	Y
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.	Y
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*

Detailed Evaluation Work Plan

For convenience, the table listing the components and the technical partners involved is shown here as a reference.

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

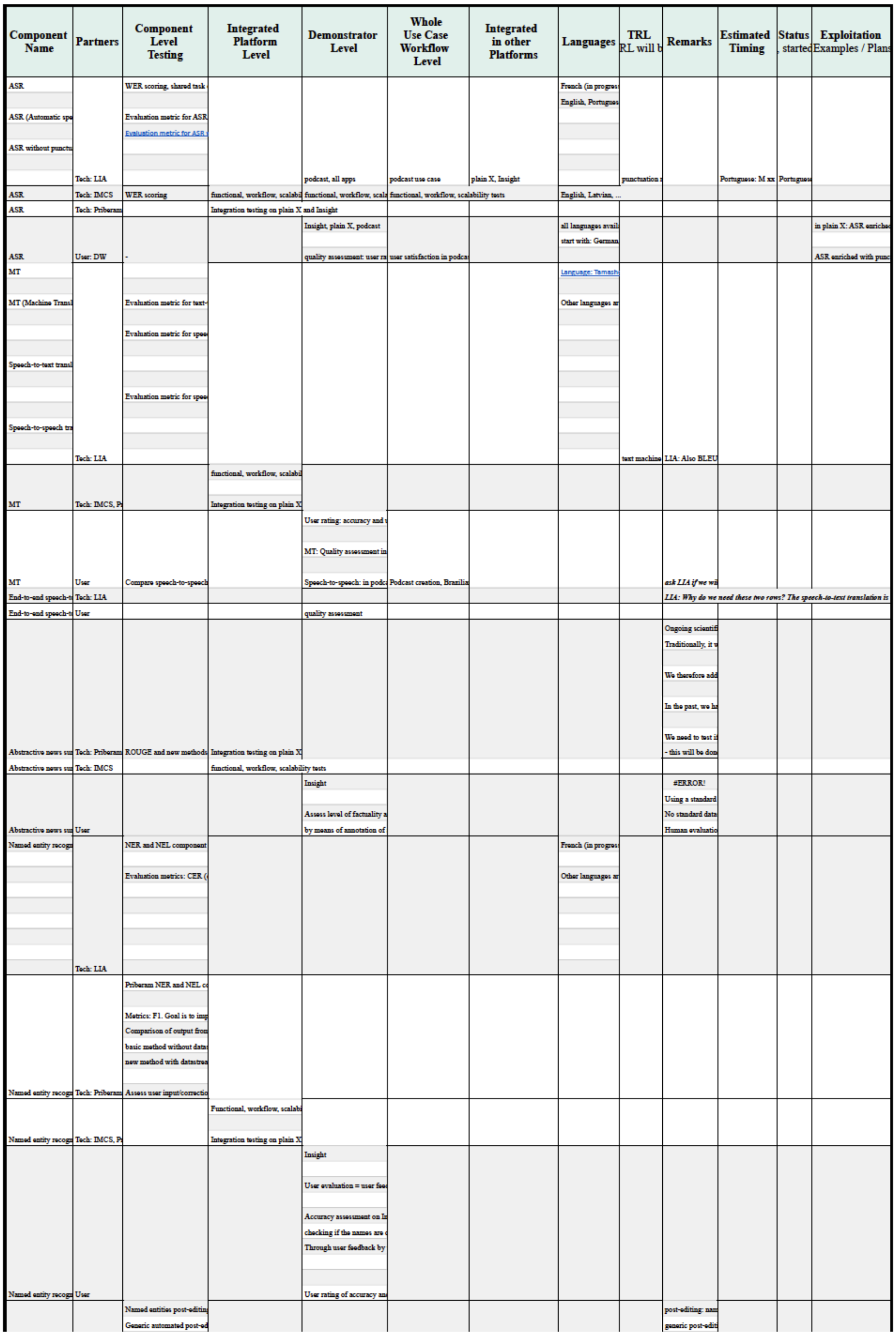
4. Named Entity Recognition (NER), Named Entity Linking (NEL), discovery	LIA, Priberam, ICMS, DW/users
5. Automatic post-editing of transcriptions and translations	LIA, FhG, Priberam, IMCS, DW/users
6. News clustering	Priberam, FhG, IMCS, DW/users
7. Topic detection	Priberam, FhG, IMCS, DW/users
8. Speech 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*

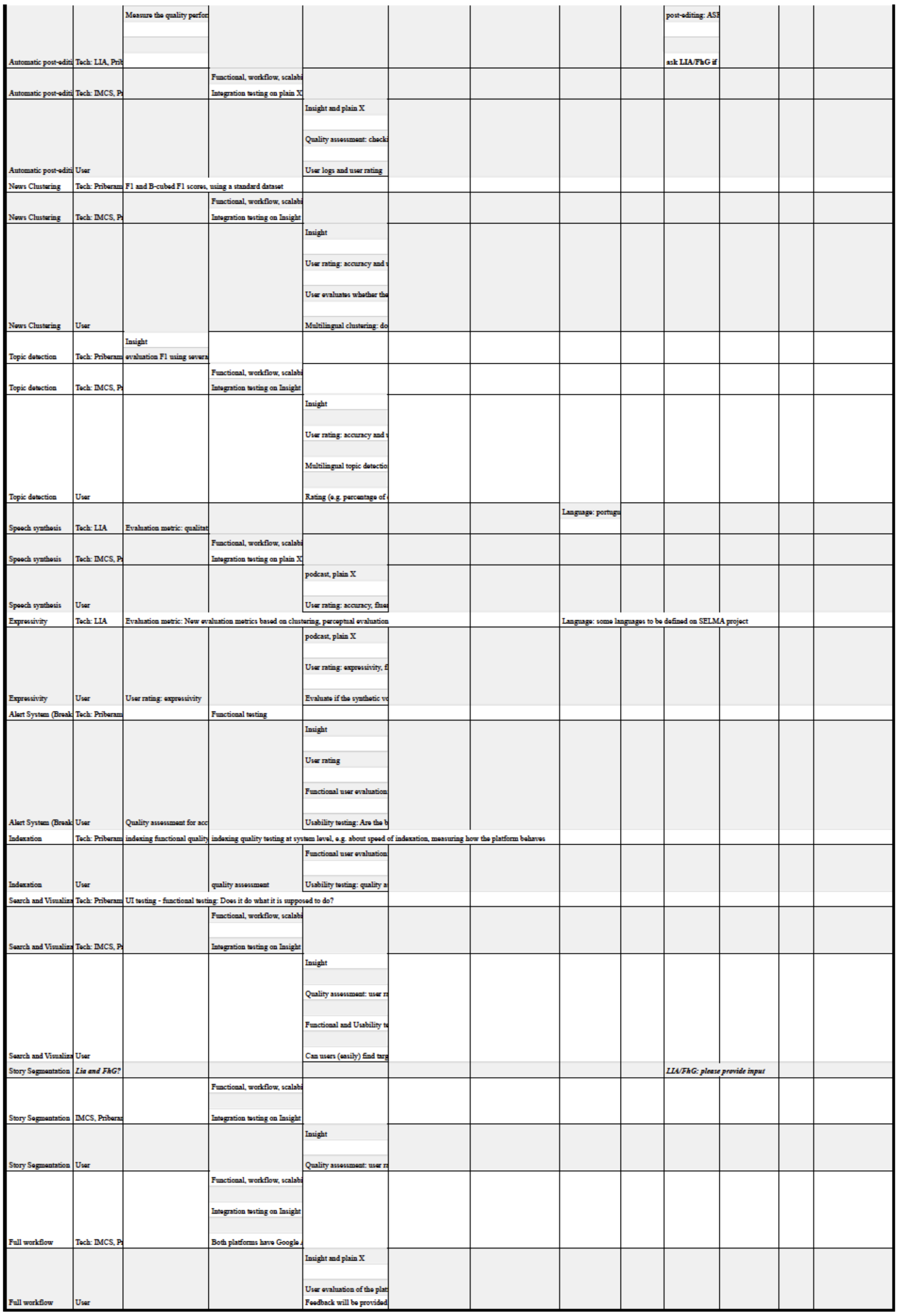
Detailed Component Evaluation Tracker

The Evaluation Excel Sheet below shows an updated status and keeps track of the evaluation activities at the different levels and by the different consortium partners. This table serves as our main evaluation 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. It is a live document and is continuously updated and expanded throughout the project.

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



Component Name	Partners	Component Level Testing	Integrated Platform Level	Demonstrator Level	Whole Use Case Workflow Level	Integrated in other Platforms	Languages	TRL RL will b	Remarks	Estimated Timing	Status , started	Exploitation Examples / Plans
ASR ASR (Automatic spe ASR without punctu	Tech: LIA	WER scoring, shared task Evaluation metric for ASR Evaluation metric for ASR		podcast, all apps	podcast use case	plain X, Insight	French (in progres English, Portugues	punctuation		Portuguese: M xx	Portugues	
ASR	Tech: IMCS	WER scoring	functional, workflow, scalabil	functional, workflow, scala	functional, workflow, scalability tests		English, Latvian, ...					
ASR	Tech: Priberam		Integration testing on plain X and Insight									
ASR	User: DW	-		Insight, plain X, podcast quality assessment: user ra	user satisfaction in podcas		all languages avail start with: German					in plain X: ASR enriche ASR enriched with punc
MT MT (Machine Transl Speech-to-text transl Speech-to-speech tr	Tech: LIA	Evaluation metric for text- Evaluation metric for spee Evaluation metric for spee					Language: Tamash Other languages ar	text machine	LIA: Also BLEU			
MT	Tech: IMCS, Pr		functional, workflow, scalabil Integration testing on plain X									
MT	User	Compare speech-to-speech		User rating: accuracy and MT: Quality assessment in Speech-to-speech: in podc	Podcast creation, Brazilia				*ask LIA if we wi*			
End-to-end speech-t	Tech: LIA								*LIA: Why do we need these two rows? The speech-to-text translation is*			
End-to-end speech-t	User			quality assessment								
Abstractive news su	Tech: Priberam	ROUGE and new methods	Integration testing on plain X						Ongoing scientifi Traditionally, it w We therefore add In the past, we ha We need to test if - this will be don			
Abstractive news su	Tech: IMCS		functional, workflow, scalability tests									
Abstractive news su	User			Insight Assess level of factuality a by means of annotation of					#ERROR! Using a standard No standard data Human evaluatio			
Named entity recogn	Tech: LIA	NER and NEL component Evaluation metrics: CER (					French (in progres Other languages ar					
Named entity recogn	Tech: Priberam	Priberam NER and NEL co Metrics: F1. Goal is to imp Comparison of output from basic method without datas new method with datastrea Assess user input/correctio										
Named entity recogn	Tech: IMCS, Pr		Functional, workflow, scalabi Integration testing on plain X									
Named entity recogn	User			Insight User evaluation = user fee Accuracy assessment on In checking if the names are Through user feedback by User rating of accuracy an								
		Named entities post-editin Generic automated post-ed							post-editing: nam generic post-editi			



Automatic post-editi	Tech: LIA, Prib	Measure the quality perfor							post-editing: AS ask LIA/FhG if			
Automatic post-editi	Tech: IMCS, Pr		Functional, workflow, scalabi Integration testing on plain X									
Automatic post-editi	User			Insight and plain X Quality assessment: checki User logs and user rating								
News Clustering	Tech: Priberam	F1 and B-cubed F1 scores, using a standard dataset										
News Clustering	Tech: IMCS, Pr		Functional, workflow, scalabi Integration testing on Insight									
News Clustering	User			Insight User rating: accuracy and u User evaluates whether the Multilingual clustering: do								
Topic detection	Tech: Priberam	Insight evaluation F1 using severa										
Topic detection	Tech: IMCS, Pr		Functional, workflow, scalabi Integration testing on Insight									
Topic detection	User			Insight User rating: accuracy and u Multilingual topic detectio Rating (e.g. percentage of								
Speech synthesis	Tech: LIA	Evaluation metric: qualitat					Language: portugu					
Speech synthesis	Tech: IMCS, Pr		Functional, workflow, scalabi Integration testing on plain X									
Speech synthesis	User			podcast, plain X User rating: accuracy, fluen								
Expressivity	Tech: LIA	Evaluation metric: New evaluation metrics based on clustering, perceptual evaluation					Language: some languages to be defined on SELMA project					
Expressivity	User	User rating: expressivity		podcast, plain X User rating: expressivity, f Evaluate if the synthetic vo								
Alert System (Break	Tech: Priberam		Functional testing									
Alert System (Break	User	Quality assessment for acc		Insight User rating Functional user evaluation Usability testing: Are the b								
Indexation	Tech: Priberam	indexing functional quality	indexing quality testing at system level, e.g. about speed of indexation, measuring how the platform behaves									
Indexation	User		quality assessment	Functional user evaluation Usability testing: quality a								
Search and Visualiza	Tech: Priberam	UI testing - functional testing: Does it do what it is supposed to do?										
Search and Visualiza	Tech: IMCS, Pr		Functional, workflow, scalabi Integration testing on Insight									
Search and Visualiza	User			Insight Quality assessment: user ra Functional and Usability te Can users (easily) find targ								
Story Segmentation	*Lia and FhG?*								*LIA/FhG: please provide input*			
Story Segmentation	IMCS, Priberan		Functional, workflow, scalabi Integration testing on Insight									
Story Segmentation	User			Insight Quality assessment: user ra								
Full workflow	Tech: IMCS, Pr		Functional, workflow, scalabi Integration testing on Insight Both platforms have Google A									
Full workflow	User			Insight and plain X User evaluation of the plat Feedback will be provided								

Figure 1 Screenshot of Updated Excel Sheet with Detailed Evaluation Plan

Test Users and User Group

The test user group contains primarily DW media professionals, selected based on their language proficiency and use case interest. Thus, specific groups have been formed for NLP benchmarking, media monitoring, news production, diversity analysis, podcasting and speech synthesis, for instance. We involved users from the Arabic, Serbian, Turkish, Brazilian, Indonesian, Kiswahili departments, as well as the Archive and Documentation Center, the Business Department and the Technology Strategy Department. In addition, a test team consisting of Priberam clients, such as EMBRAER, AICEP and LUSA, are also evaluating the SELMA enhancements to the integrated platforms Monitio and plain X.

The project has also set up a User Group of 31members, from the EBU, SWR, ARTE, RAI, BBC, EuroNews, Prisa, EMBRAER, AICEP, and some research organizations such as the University of Edinburgh and the University of Tilburg. The members of the User Group and Advisory Board, a subset of the User Group, are asked to provide feedback and recommendations and support further outreach.

SELMA had its first User Day at Deutsche Welle in Bonn on 12 October 2022. This hybrid event included presentations of the SELMA progress over the first project half. We had up to 60 participants in the remote sessions (presentations and panel discussion) and 20 in the onsite workshop and demo sessions in the afternoon. The remote sessions allowed members of the audience, including those from the official SELMA User Group, to comment, ask questions, and provide suggestions. This opened up the view and allowed for some new directions. More details on the SELMA user day can be found in D6.4 - Interim Impact Report.

3. Technical Testing

This section addresses technical testing for individual components as well as integrated platforms and demonstrators. We mainly list the components and platforms that are subject to technical testing and add details as to the current status to the Evaluation Tracking Document, the updated Excel sheet.

Specific technical details are reported in the respective technical deliverables:

- D1.3 Intermediate prototype report
- D2.4 Intermediate progress report on continuous massive stream learning
- D2.5 Intermediate release of stream learning and entity linking capabilities
- D2.6 Intermediate release of segmentation, summarization and news classification capabilities
- D3.4 Intermediate progress report on speech and natural language processing
- D3.5 Intermediate release of transcription, punctuation and translation, voice synthesis capabilities
- D3.6 Intermediate release of post-editing and user feedback capabilities
- D4.3 Intermediate platform with continuous massive stream learning NLP capabilities focuses on the integration efforts.

SELMA NLP components developed by the University of Avignon (LIA), Fraunhofer (FhG), Priberam, and IMCS are primarily 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.

Evaluation of the 12 components listed in Table 2 - Basic Component Overview started in Year 1, and intensified in the second project year. The evaluation of most of the components is currently underway, one (alert system) will commence in the final year.

- Component 1: ASR – ongoing – see D3.4 Intermediate progress report on speech and natural language processing; D3.5 Intermediate release of transcription, punctuation and translation, voice synthesis capabilities

- Component 2: MT – ongoing - D3.4 Intermediate progress report on speech and natural language processing; D3.5 Intermediate release of transcription, punctuation and translation, voice synthesis capabilities
- Component 3: Abstractive Summarization – ongoing – see D2.4 Intermediate progress report on continuous massive stream learning; D2.6 Intermediate release of segmentation, summarization and news classification capabilities
- Component 4: NER/NEL - ongoing – see D2.4 Intermediate progress report on continuous massive stream learning; D2.5 Intermediate release of stream learning and entity linking capabilities; D3.4 Intermediate progress report on speech and natural language processing; D3.5 Intermediate release of transcription, punctuation and translation, voice synthesis capabilities
- Component 5: Automatic post-editing of ASR and MT – ongoing – see D3.6 Intermediate release of post-editing and user feedback capabilities
- Component 6: News Clustering – ongoing - see D2.4 Intermediate progress report on continuous massive stream learning; D2.6 Intermediate release of segmentation, summarization and news classification capabilities
- Component 7: Topic Detection – ongoing – see D2.4 Intermediate progress report on continuous massive stream learning; D2.6 Intermediate release of segmentation, summarization and news classification capabilities
- Component 8: Speech Synthesis – ongoing – see D3.4 Intermediate progress report on speech and natural language processing; D3.5 Intermediate release of transcription, punctuation and translation, voice synthesis capabilities
- Component 9: Alert System – planned for Y3
- Component 10: Indexation – ongoing – see D2.4 Intermediate progress report on continuous massive stream learning; D2.6 Intermediate release of segmentation, summarization and news classification capabilities
- Component 11: Search and Visualization in UI – ongoing - D1.3 Intermediate prototype report
- Component 12: Story Segmentation – ongoing – see D2.4 Intermediate progress report on continuous massive stream learning; D2.6 Intermediate release of segmentation, summarization and news classification capabilities

- Integrated SELMA OSS – ongoing – see D1.3 Intermediate prototype report; D4.3 Intermediate platform with continuous massive stream learning NLP capabilities focuses on the integration efforts
- Integrated plain X platform – ongoing – see D1.3 Intermediate prototype report; D4.3 Intermediate platform with continuous massive stream learning NLP capabilities focuses on the integration efforts
- Integrated Monitio platform – ongoing – see D1.3 Intermediate prototype report; D4.3 Intermediate platform with continuous massive stream learning NLP capabilities focuses on the integration efforts

An overview of the current status per component can be found in more detail in the table featured in Figure 1 - Detailed Component Evaluation Tracker.

4. User Evaluation

User Evaluation has taken place at several levels. We evaluate the individual components in close cooperation with the developers, at platform level (SELMA OSS, Monitio and plain X) with new SELMA components integrated through their new or enhanced functionalities, as well as through the use cases and use case applications.

Although Usability testing - pursuing the five E's: effective, efficient, engaging, error-tolerant and easy to learn (https://www.wqusability.com/) - will be the focus of Y3, this has already been done to some degree in these earlier phases through user evaluation at platform and use case level. This is the case for the SELMA OSS, Monitio and plain X platforms, where the new SELMA features and functionalities have been tried out, and for the podcasting and diversity use cases evaluated at use case level. In Y3, concrete and measurable feedback on usability will be obtained through user observation, questionnaires, and interviews.

4.1 Integrated SELMA OSS

In **Year 1**, it was decided to add an extra use case, i.e., UC0, enabling SELMA's open-source software (OSS) components to be made accessible via a unified API under the umbrella of use case 0. Several UIs and tools were made available through the OSS platform and initial testing of those UIs and tools was done in the first year.

This API offers the possibility to evaluate individual components through command-line tools and more advanced user applications.

In **Year 2,** improvements on the OSS tools were further evaluated, including the basic transcription-translation-speech synthesis workflow in this UI. New NLP engines were added in Y2, including the speech translation developed by LIA and the DW customized Brazilian voices. This workflow was tested from a user point of view, in terms of ease of use, easy access, speed and language coverage. It is considered useful as a low-threshold tool for demoing this kind of workflow and a basic view of automation of NLP processes, but is limited for choice of engines (only OSS engines). That is acceptable because of the purpose of the OSS, while a more sophisticated platform with a variety of additional commercial engines is available through plain X.

In addition, two components of the SELMA OSS were evaluated for user applications using the command-line API.

First, low-resourced language translation modules from the GoURMET – Global Under-Resourced Media Translation – a Horizon 2020 project – in which Deutsche Welle participated and which ended in March 2022 - were made available on-demand through SELMA's DockerSpaces orchestrator. This enabled DW as end user to try out and put the MT engines coming from this EU project (Grant agreement 825299) to use. All 16 engines were integrated into the SELMA OSS. The engines were installed in SELMA as dockerized modules.

A command-line tool allowed us to put the system through its paces by sending repeated translation requests for a variety of source and target languages. The following figure shows the number of words per seconds for various source/target language pairs that the system managed to translate. Running the script repeatedly allowed us to test the orchestrator's stability and dependability, while highlighting areas that required improvements.

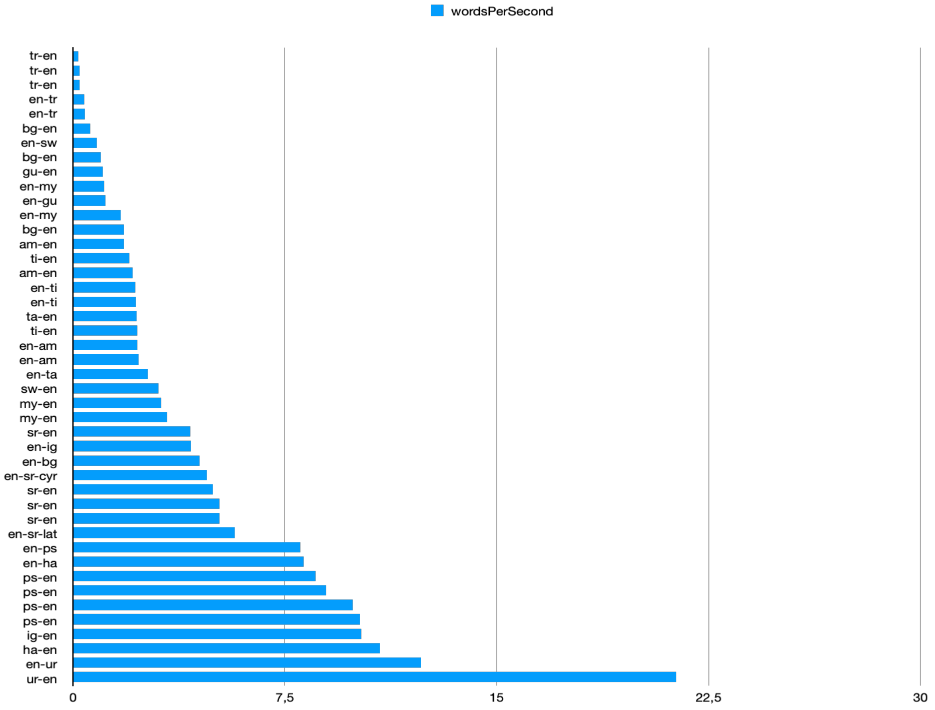


Figure 2 *Benchmarking of GoURMET modules with SELMA OSS*

The GoURMET evaluation was extended with a Word Cloud analysis application. This iOS application was developed by the DW SELMA project team. It starts with a batch translation of a selected set of texts, translated by selected engines, in this case the GoURMET dockerized modules in SELMA OSS. The translations were automatically retrieved and a subsequent analysis for content focus was done, resulting in a Word Cloud output. This allows us to do automated batch processing of selected source documents for a topical content analysis.

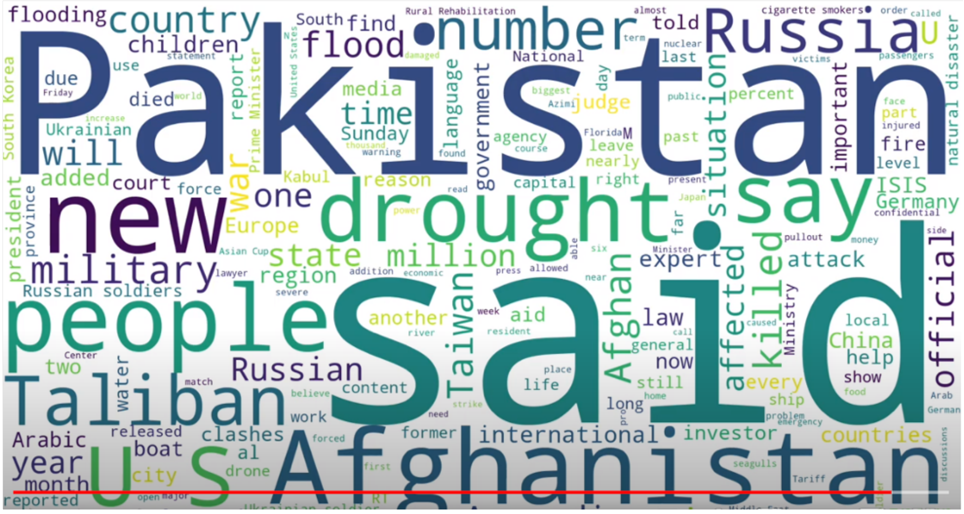


Figure 3 Word Cloud application in SELMA OSS

The following screencast shows how the Word Cloud application works:

https://www.youtube.com/watch?v=ejNTUd9Tda0

Second, the Podcast creator application uses a similar DockerSpaces API to convert news bulletin texts to Brazilian speech. Again, this allowed us to highlight areas where either the text-to-speech Docker module needed updating or the Orchestrator's stability required improvements. More details on the Podcast creator evaluation are given in section 4.5.

In **Year 3**, we will continue to try out the orchestration in UC0 and the SELMA Dockerspaces, do comparative analysis and look for additional applications.

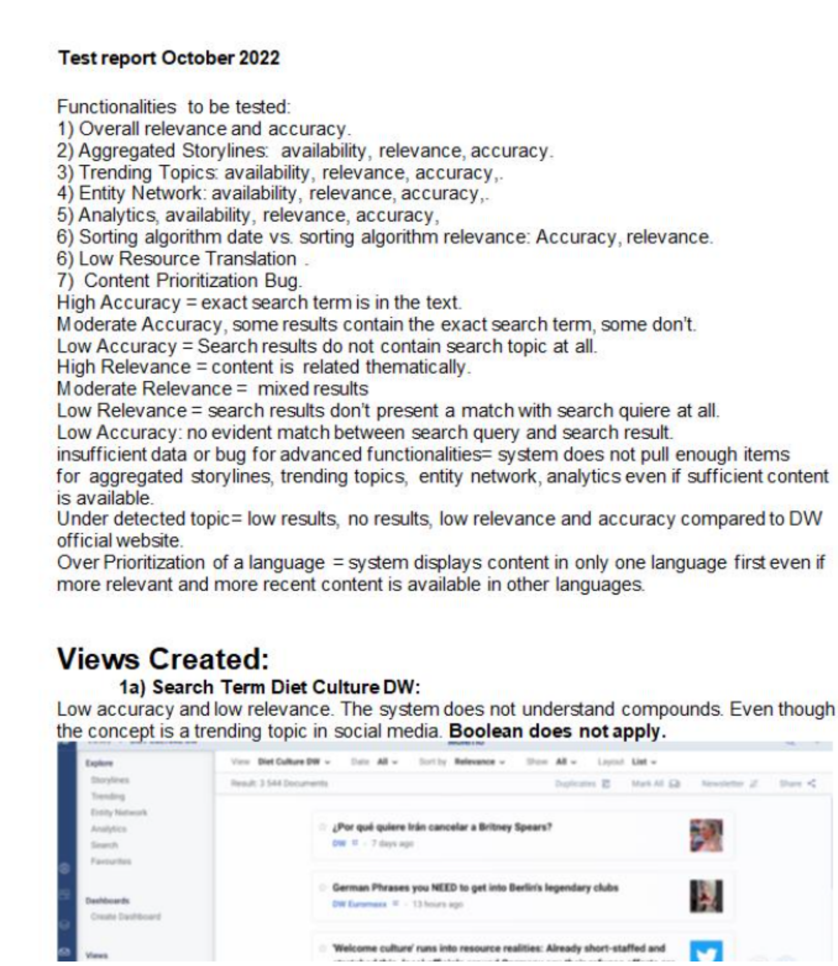
4.2 Monitio Demonstrator

The Monitio platform is the main demonstrator for the first use case (UC1), multilingual media monitoring, as described in D1.1 - Use Case Description and Requirements. It scans a wide collection of media items and provides a sophisticated filtering and automatic analysis system, producing a selection of clustered news items by topic or another common attribute, according to the user's preferences.

In **Year 1**, the platform was introduced to the consortium and requirements and user scenarios were set. It was decided which NLP enhancements to the tool were needed and were within the scope of the project and the consortium. In particular improvements as to integration, analysis speed, named entity recognition and linking, building dictionaries and integrating thesauri, incorporating user feedback mechanisms, and user interfacing were addressed.

User evaluation will evaluate each of the added modules and features as well as the overall usability of the platform, in terms of functionality, accuracy, usefulness and acceptance for media monitoring and decision making.

Throughout **Year 2**, Deutsche Welle tried out the Monitio platform and provided feedback on the overall UI, functionalities, ease of use, transparency and consistency. We discussed possible use cases internally with media professionals, setting priorities and suggesting changes. Content feeds were provided by Deutsche Welle and integrated into the platform. After numerous trials, it was decided that the newsletter production is a prime goal for wider distribution and translation of the headlines and cluster summarizations into the user's preferred target language. Detailed input on functionalities such as filtering, search parameters, language settings, was given. The focus on text analysis makes senses, but the content ingestion should be expanded to include video and audio formats. YouTube videos were added in this reporting period on user demand. A renewed UI was developed towards the end of the year and testing is currently underway, with regular feedback.



Test report October 2022

Functionalities to be tested:
1) Overall relevance and accuracy.
2) Aggregated Storylines: availability, relevance, accuracy.
3) Trending Topics: availability, relevance, accuracy,.
4) Entity Network: availability, relevance, accuracy,.
5) Analytics, availability, relevance, accuracy,
6) Sorting algorithm date vs. sorting algorithm relevance: Accuracy, relevance.
6) Low Resource Translation .
7) Content Prioritization Bug.
High Accuracy = exact search term is in the text.
Moderate Accuracy, some results contain the exact search term, some don't.
Low Accuracy = Search results do not contain search topic at all.
High Relevance = content is related thematically.
Moderate Relevance = mixed results
Low Relevance = search results don't present a match with search quiere at all.
Low Accuracy: no evident match between search query and search result.
insufficient data or bug for advanced functionalities= system does not pull enough items for aggregated storylines, trending topics, entity network, analytics even if sufficient content is available.
Under detected topic= low results, no results, low relevance and accuracy compared to DW official website.
Over Prioritization of a language = system displays content in only one language first even if more relevant and more recent content is available in other languages.

Views Created:

1a) Search Term Diet Culture DW:

Low accuracy and low relevance. The system does not understand compounds. Even though the concept is a trending topic in social media. **Boolean does not apply.**

View: Diet Culture DW · Date: All · Sort by: Relevance · Show: All · Layout: List

Result: 3 544 Documents

¿Por qué quiere Irán cancelar a Britney Spears?
DW · 7 days ago

German Phrases you NEED to get into Berlin's legendary clubs
DW Euromaxx · 13 hours ago

'Welcome culture' runs into resource realities: Already short-staffed and

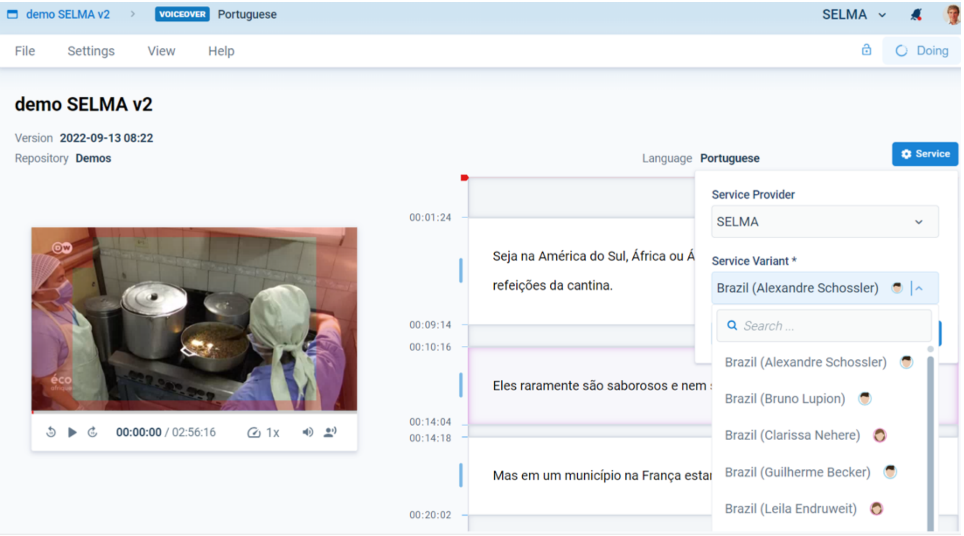
Figure 3 Example of feedback on the Monitio functionalities

Year 3 will see a thorough and wide user evaluation of the updated/finalized platform, with different departments, including Archive and Documentation for specialized searching, retrieval and integration, and editorial language departments for internal monitoring of DW content and creation and use of customized Monitio Newsletters.

4.3 plain X Demonstrator

In **Year 1**, we established the requirements, scenarios and evaluation plan for the content creation use case (UC2), with the plain X demonstrator as the prime target platform. In this period, we determined the initial status of the plain X platform, which has been developed over several years and is a platform which is being further jointly developed and exploited by Priberam and DW. It targets a smooth workflow for multilingual transcription, translation, subtitling and voice-over with synthetic voices. User requirements, workflows and enhancements that can be developed within SELMA were discussed.

Year 2 focused on evaluating envisaged integrated enhancements, including ASR modules, speech-to-translated-text and customized synthetic voices for Brazilian Portuguese, all modules from the University of Avignon. End users started evaluating the output and comparing it with other processes and tools. This is described in sections 4.7, 4.8, and 4.9 of this report.



demo SELMA v2

Version 2022-09-13 08:22
Repository Demos

Language Portuguese

Service Provider: SELMA

Service Variant *: Brazil (Alexandre Schossler)

Search ...

- Brazil (Alexandre Schossler)
- Brazil (Bruno Lupion)
- Brazil (Clarissa Nehere)
- Brazil (Guilherme Becker)
- Brazil (Leila Endruweit)

00:01:24

Seja na América do Sul, África ou Á
refeições da cantina.

00:09:14

00:10:16

Eles raramente são saborosos e nem

00:14:04

00:14:18

Mas em um município na França estar

00:20:02

***Figure 4** Brazilian customized voices in plain X*

In **Year 3**, we will do an in-depth comparative automated and user analysis of the different processes, including the traditional ASR + MT workflow vs. speech-to-translated text, as well as ASR + MT + voice-over vs. speech-to-speech translation. The plain X platform is an ideal environment for such comparative evaluation.

We also plan to perform an assessment of additional SELMA components to be developed or enhanced in the final year, including a punctuation module for some languages that do not have punctuation in ASR, improved named entity recognition, a built-in feedback mechanism, and new or improved synthetic voices. We also expect that speaker diarization in the ASR mode and voice selection based on diarization in the voice-over mode is part of the final evaluation phases.

4.4 Diversity Use Case Application

We established and described the Diversity Use Case Application in **Year 1.** This is an application of the news monitoring use case and assesses the ability of the Monitio platform to analyze an arbitrary group of articles with respect to the diversity of their content. We agreed upon the metadata that needed to be added and the source and limitations, i.e., use of only Wikidata for entities and their added metadata, and a suggestion of fields to be added in the UI.

In **Year 2**, this materialized in the form of adding a diversity category in the platform, for named entities, based on metadata from Wikidata. 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. These fields can be used for filtering and viewing available information. The intended output is statistics on gender (and other minority groups) balance.

The evaluation focused on obtaining accurate and useful statistics. Detailed searches revealed the level of information that can be obtained on these categories and triggered a discussion of what is ethical and permitted in this respect and what the risks are in case of unintended use.

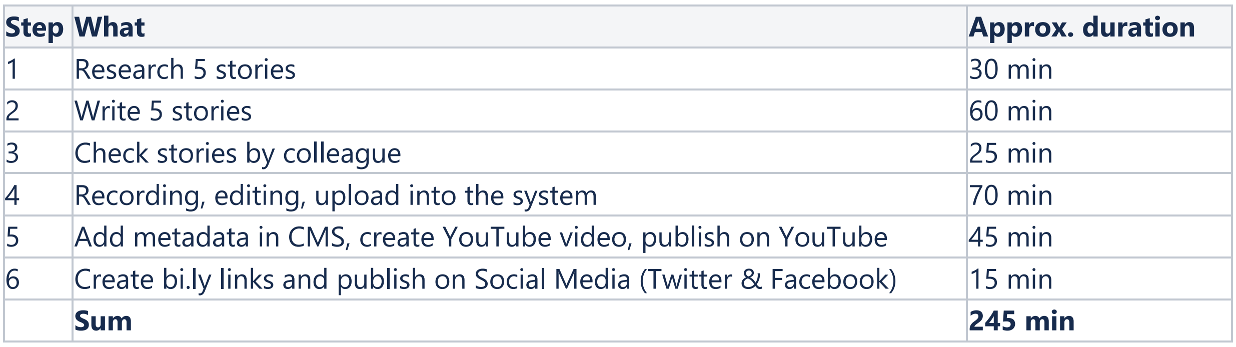
Year 3 will focus on the ethical consequences in the first place, with the necessary adaptations of the use of the categories and filtering options. Further evaluations will see in how far the tool provides useful and reliable statistical data on gender and minority representation, for instance in the pursuit of the 50/50 gender balance project involving DW and other broadcasters.

4.5 Podcast Use Case Application

The Podcast Creator use case is based on a workflow observed in DW's Brazilian language department. The use case's goal is to increase the workflow's efficiency by supporting the journalist in the production of daily audio news bulletins through SELMA.

The production of a single news bulletin can be subdivided into the following steps. The table shows the duration that is required for each step during the classic, manual process.

In **Year 1,** we established the concept for this use case and initiated contact with editorial users that could be involved and set requirements.



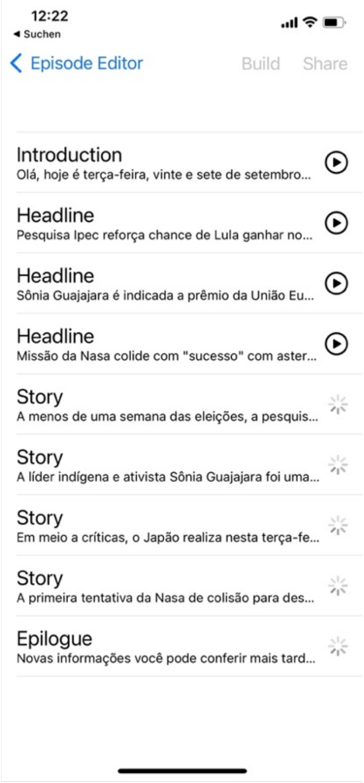
Step	What	Approx. duration
1	Research 5 stories	30 min
2	Write 5 stories	60 min
3	Check stories by colleague	25 min
4	Recording, editing, upload into the system	70 min
5	Add metadata in CMS, create YouTube video, publish on YouTube	45 min
6	Create bi.ly links and publish on Social Media (Twitter & Facebook)	15 min
	Sum	**245 min**

Figure 5 *Traditional Process for Podcasting Use Case*

The work in **Year 2** consisted in creating and trying out an automated process to speed up and facilitate some of the above steps. It uses speech technology in the form of DW customized voices developed by LIA, made available through integration with the SELMA OSS, and a newly created Podcast iOS app, This work involved technology partners, such as IMCS and LIA, as well as DW project managers and editors from the Brazilian department, native speakers used to doing the work in the traditional way. These news stories are produced and published twice a day. The SELMA enhanced module includes a template that streamlines the production process for editorial users and automatically inserts basic, recurring components such as the introduction and music in between the stories.

For this use case we also assessed the quality of the Brazilian customized voices. More details on this are available in section 4.9.

The application already automates the task of creating speech and mixing it with music and background tracks. The following screenshot shows the app while synthesizing speech for the various sections of a Brazilian news podcast by accessing the API provided through UC0.



12:22
◂ Suchen

Introduction
Olá, hoje é terça-feira, vinte e sete de setembro...

Headline
Pesquisa Ipec reforça chance de Lula ganhar no...

Headline
Sônia Guajajara é indicada a prêmio da União Eu...

Headline
Missão da Nasa colide com "sucesso" com aster...

Story
A menos de uma semana das eleições, a pesquis...

Story
A líder indígena e ativista Sônia Guajajara foi uma...

Story
Em meio a críticas, o Japão realiza nesta terça-fe...

Story
A primeira tentativa da Nasa de colisão para des...

Epilogue
Novas informações você pode conferir mais tard...

Figure 6 *SELMA Podcast Use Case Automation Template*

***Figure** 7 Demonstrating the Podcast Use Case at the SELMA user day in Bonn*

The semi-automated process and template was also demonstrated at the SELMA user day workshop in October 2022. Two editorial DW departments, in particular the Hindi and the Urdu departments tried it out, compared it with their current process, and expressed interest in participating in future user trials.

The plans for **Year 3** are to involve editorial teams to further try out the automation template in real production, to assess it in terms of gain in effort and productivity. Thus, once the Podcast creator is in real use, a good benchmark will be to do these measurements again and compare them with the original production process shown above.

Additional improvements are planned, in particular in step number 4, recording, editing and uploading into the system.

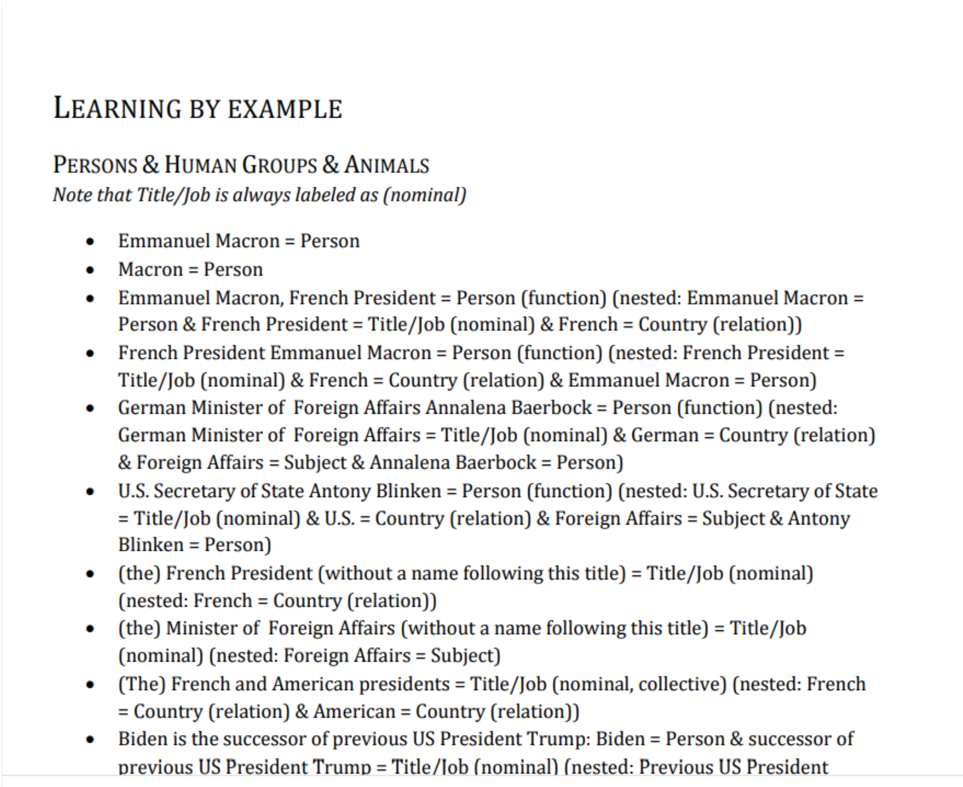
More improvements should be observable for steps 1 (as an integration with the Monitio monitoring platform is foreseen, resulting in an automation of pre-selecting suitable stories) and step 2 (as SELMA's summarization component should provide the journalist with a good starting point for adapting a news story to its audio version).

In the end evaluation, we attempt to reveal where the main benefits are: using the app with a template, using synthetic voices instead of speakers, the entire automated workflow compared to the manual one, having the editor or the monitoring platform select the items, having the editor write the stories or using and editing the SELMA summarization component.

4.6 NER component

Year 1 determined the process for the technical as well as the user partners in terms of developing and training the named entity recognition component. It was agreed that certain languages will be targeted, and annotation should be done by native speakers. Training of the tool with previously created datasets was initiated. Annotation of some additional languages was started, including Ukrainian, Latvian and Russian. Training of DW project managers was started for Arabic.

In Year 2, the annotation process was streamlined and adapted, based on experiences on the first language set. The initial requirement to obtain 4,000 documents in each annotation language was revisited and reduced to 50-500, depending on the quality of the pre-annotation. This was necessary to keep the effort required for this task manageable and reasonable. A generic, multilingual annotator was built based on the first annotation sets, which allows for a pre-annotation of datasets in additional languages, thus reducing the need for such a high level of human annotation. The final human annotation level differs per language, for instance for Dutch, an annotation of 50 documents was sufficient for a very good result. Turkish, on the other hand, does not reveal such good results and needs more annotation. Deutsche Welle intensified the training and preparation and set up a detailed information package for the editors that will be involved in the annotation. This includes a special DW user guide for editors, with selected and sorted examples, to make the introduction as smooth as possible. Detailed feedback on the initial guidelines and the UI was provided to the annotation linguists. In this reporting period, the partners completed Latvian and Dutch, and started Russian, Ukrainian, Turkish, Arabic and Urdu.



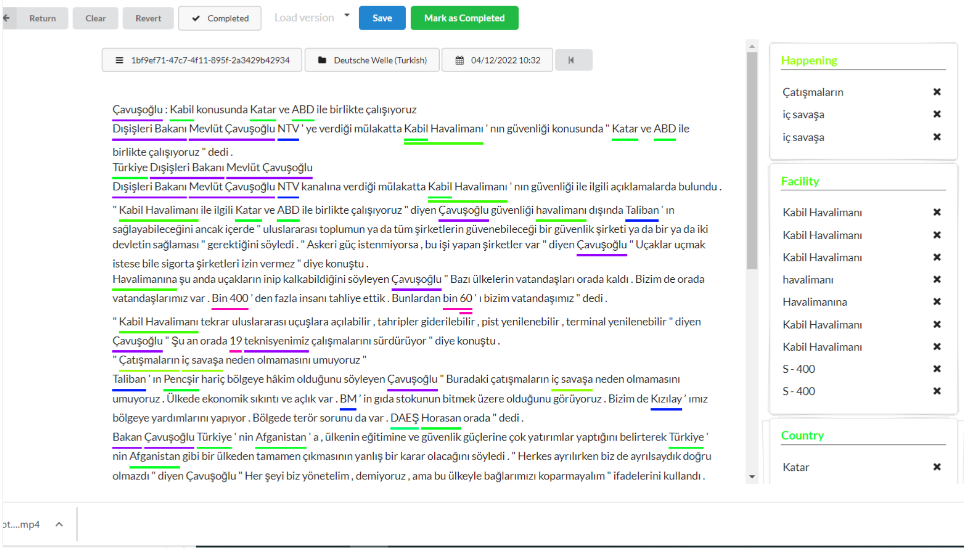
LEARNING BY EXAMPLE

PERSONS & HUMAN GROUPS & ANIMALS

Note that Title/Job is always labeled as (nominal)

- Emmanuel Macron = Person
- Macron = Person
- Emmanuel Macron, French President = Person (function) (nested: Emmanuel Macron = Person & French President = Title/Job (nominal) & French = Country (relation))
- French President Emmanuel Macron = Person (function) (nested: French President = Title/Job (nominal) & French = Country (relation) & Emmanuel Macron = Person)
- German Minister of Foreign Affairs Annalena Baerbock = Person (function) (nested: German Minister of Foreign Affairs = Title/Job (nominal) & German = Country (relation) & Foreign Affairs = Subject & Annalena Baerbock = Person)
- U.S. Secretary of State Antony Blinken = Person (function) (nested: U.S. Secretary of State = Title/Job (nominal) & U.S. = Country (relation) & Foreign Affairs = Subject & Antony Blinken = Person)
- (the) French President (without a name following this title) = Title/Job (nominal) (nested: French = Country (relation))
- (the) Minister of Foreign Affairs (without a name following this title) = Title/Job (nominal) (nested: Foreign Affairs = Subject)
- (The) French and American presidents = Title/Job (nominal, collective) (nested: French = Country (relation) & American = Country (relation))
- Biden is the successor of previous US President Trump: Biden = Person & successor of previous US President Trump = Title/Job (nominal) (nested: Previous US President

Figure 8 *Customized DW Guidelines for NER Annotation Editors*



1bf9ef71-47c7-4f11-895f-2a3429b42934 | Deutsche Welle (Turkish) | 04/12/2022 10:32

Çavuşoğlu : Kabil konusunda Katar ve ABD ile birlikte çalışıyoruz
Dışişleri Bakanı Mevlüt Çavuşoğlu NTV ' ye verdiği mülakatta Kabil Havalimanı ' nın güvenliği konusunda " Katar ve ABD ile birlikte çalışıyoruz " dedi .
Türkiye Dışişleri Bakanı Mevlüt Çavuşoğlu
Dışişleri Bakanı Mevlüt Çavuşoğlu NTV kanalına verdiği mülakatta Kabil Havalimanı ' nın güvenliği ile ilgili açıklamalarda bulundu .
" Kabil Havalimanı ile ilgili Katar ve ABD ile birlikte çalışıyoruz " diyen Çavuşoğlu güvenliği havalimanı dışında Taliban ' ın sağlayabileceğini ancak içerde " uluslararası toplumun ya da tüm şirketlerin güvenebileceği bir güvenlik şirketi ya da bir ya da iki devletin sağlaması " gerektiğini söyledi . " Askeri güç istenmiyorsa , bu işi yapan şirketler var " diyen Çavuşoğlu " Uçaklar uçmak istese bile sigorta şirketleri izin vermez " diye konuştu .
Havalimanına şu anda uçakların inip kalkabildiğini söyleyen Çavuşoğlu " Bazı ülkelerin vatandaşları orada kaldı . Bizim de orada vatandaşlarımız var . Bin 400 ' den fazla insanı tahliye ettik . Bunlardan bin 60 ' ı bizim vatandaşımız " dedi .
" Kabil Havalimanı tekrar uluslararası uçuşlara açılabilir , tahripler giderilebilir , pist yenilenebilir , terminal yenilenebilir " diyen Çavuşoğlu " Şu an orada 19 teknisyenimiz çalışmalarını sürdürüyor " diye konuştu .
" Çatışmaların iç savaşa neden olmamasını umuyoruz "
Taliban ' ın Pençşir hariç bölgeye hâkim olduğunu söyleyen Çavuşoğlu " Buradaki çatışmaların iç savaşa neden olmamasını umuyoruz . Ülkede ekonomik sıkıntı ve açlık var . BM ' in gıda stokunun bitmek üzere olduğunu görüyoruz . Bizim de Kızılay ' ımız bölgeye yardımlarını yapıyor . Bölgede terör sorunu da var . DAEŞ Horasan orada " dedi .
Bakan Çavuşoğlu Türkiye ' nin Afganistan ' a , ülkenin eğitimine ve güvenlik güçlerine çok yatırımlar yaptığını belirterek Türkiye ' nin Afganistan gibi bir ülkeden tamamen çıkmasının yanlış bir karar olacağını söyledi . " Herkes ayrılırken biz de ayrılsaydık doğru olmazdı " diyen Çavuşoğlu " Her şeyi biz yönetelim , demiyoruz , ama bu ülkeyle bağlarımızı koparmayalım " ifadelerini kullandı .

Happening
Çatışmaların ✖
iç savaşa ✖
iç savaşa ✖

Facility
Kabil Havalimanı ✖
Kabil Havalimanı ✖
Kabil Havalimanı ✖
havalimanı ✖
Havalimanına ✖
Kabil Havalimanı ✖
Kabil Havalimanı ✖
S - 400 ✖
S - 400 ✖

Country
Katar ✖

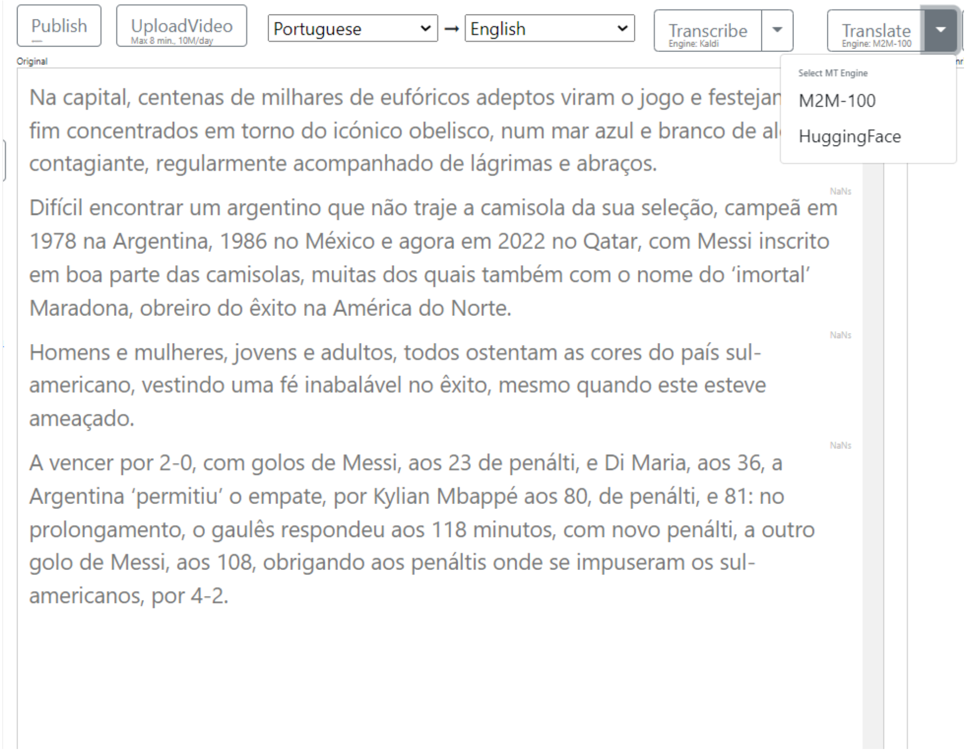
***Figure 9** Sample of Turkish NER annotation*

Year 3 is expected to see the completion of the targeted languages for the annotation dataset and a major improvement of the NER and NEL accuracy in the plain X and Monitio platforms. In-depth evaluations will be done in those two platforms to see if such improvement is visible to the end user.

4.7 MT components

Year 1 focused on setting the requirements for MT development and efforts, what is needed, what is already available, what should be integrated and what can be achieved within the project.

In terms of integration, we looked at which MT engines are/should be integrated into the different platforms in terms of efficiency and cost-effectiveness. UC0 is an open-source platform and costs should be kept to a minimum, especially if we want to make it available for wide-scale testing. M2M-100 and HuggingFace were selected as basic MT engines, as they cover a large number of languages, are fast enough and provide sufficient quality for functionality testing. The focus here is not on the translation quality, but on the workflow, processes and functionalities.



Na capital, centenas de milhares de eufóricos adeptos viram o jogo e festejar fim concentrados em torno do icónico obelisco, num mar azul e branco de al contagiante, regularmente acompanhado de lágrimas e abraços.

Difícil encontrar um argentino que não traje a camisola da sua seleção, campeã em 1978 na Argentina, 1986 no México e agora em 2022 no Qatar, com Messi inscrito em boa parte das camisolas, muitas dos quais também com o nome do 'imortal' Maradona, obreiro do êxito na América do Norte.

Homens e mulheres, jovens e adultos, todos ostentam as cores do país sul-americano, vestindo uma fé inabalável no êxito, mesmo quando este esteve ameaçado.

A vencer por 2-0, com golos de Messi, aos 23 de penálti, e Di Maria, aos 36, a Argentina 'permitiu' o empate, por Kylian Mbappé aos 80, de penálti, e 81: no prolongamento, o gaulês respondeu aos 118 minutos, com novo penálti, a outro golo de Messi, aos 108, obrigando aos penáltis onde se impuseram os sul-americanos, por 4-2.

Figure 10 *Generic MT engines in the SELMA OSS*

The plain X platform already offers a choice of MT engines, including DeepL, Google, Azure, and Facebook. Specific SELMA MT components will be added, in particular for direct speech-to-text and even speech-to-speech engines.

The Monitio platform is less focused on machine translation, but still includes it, to convert the content from different languages into the one(s) that the user has defined as the preferred language.

The overall focus of Machine Translation efforts in SELMA is on speech-to-text and speech-to-speech translation.

In **Year 2**, the speech-to-text translation module from the University of Avignon was added to the SELMA OSS and the plain X platforms. This allows a direct translation of speech within a video from English into French without going to a direct transcription.

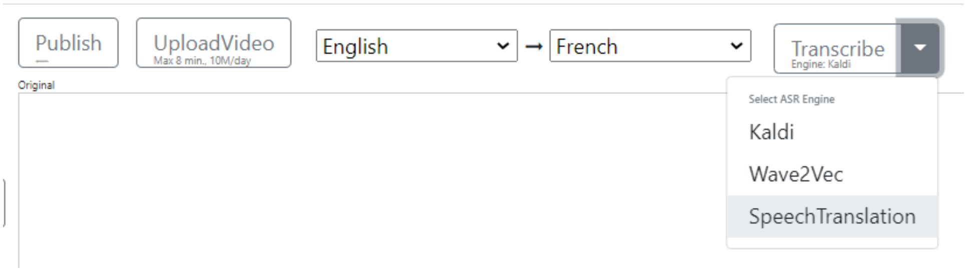


Figure 11 *ASR engines in the SELMA OSS*

This new module was evaluated by DW by testing the functionality in both platforms, to see if we actually get the French translation and how smooth the process is. At first, it worked only occasionally in plain X, due to integration issues. This feedback was forwarded to the developers and the feature was improved.

The next step was/is to compare the quality of the output of the traditional process, going through a transcription and then a translation, and the new process of going directly from speech to target language. Initial testing was done on videos from English into French translated text. This evaluation is ongoing.

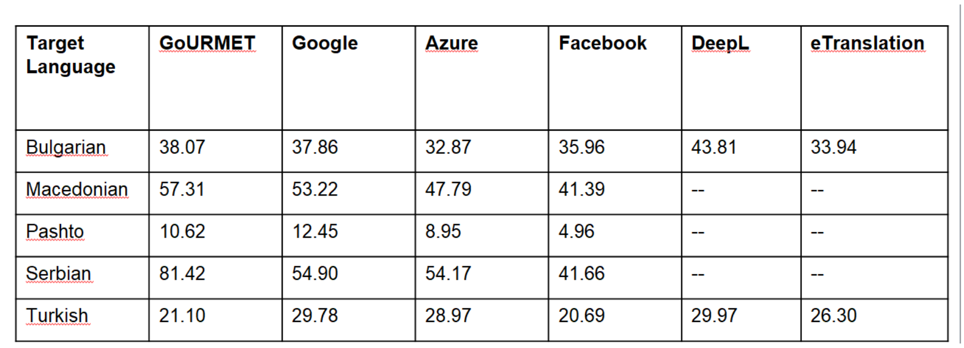
We also did an evaluation of some customized MT engines. Dockerized instances of the 32 GoURMET MT models for 16 languages (developed within the GoURMET H-2020 project) for some low-resourced languages were adapted, integrated and installed in the SELMA OSS to allow for an evaluation of the engines in the platform.

User evaluation was done at different levels:

- Texts were selected and ingested in batch into the SELMA OSS via API
- Back-to-back (reverse) translation for 16 languages was done
- A subsequent automated evaluation was done with BLEU scoring
- Reference scripts were created in different target languages
- BLEU scores were produced for languages which had a reference script
- Human evaluation was done of the best performing language pairs, based on the BLEU scores

More details of this process can be found in section 4.9 on DW NLP Benchmarking.

The table below shows a comparative analysis with BLEU scores between the available engines for the best performing GoURMET-focused low-resource languages covered by DW from back-to-back translation evaluation using the SELMA OSS. It allows us to determine the best engine for a specific language pair.



Target Language	GoURMET	Google	Azure	Facebook	DeepL	eTranslation
Bulgarian	38.07	37.86	32.87	35.96	43.81	33.94
Macedonian	57.31	53.22	47.79	41.39	--	--
Pashto	10.62	12.45	8.95	4.96	--	--
Serbian	81.42	54.90	54.17	41.66	--	--
Turkish	21.10	29.78	28.97	20.69	29.97	26.30

Figure 12 *Comparative MT evaluation in SELMA OSS of selected engines*

4.6 ASR components

Year 1 work included setting the requirements and priorities and determining what was available and what needed to be done. Both the University of Avignon already have some transcription tools, for instance for German, English, French, Arabic and Russian. LIA’s main aim is to develop a speech-to-translated-text tool and a speech-to-speech tool and incorporate the ASR into these processes. The goal of the speech-to-speech translation component that LIA is working on is to transfer human-read news segments from one language to another, while keeping the original voice’s expressivity.

FhG’s goal is to apply its ASR to live broadcasting streams, to expand it to selected low-resourced languages, and work on enhancement modules such as punctuation.

Development and technical testing was the focus in the first period.

In **Year 2**, the user evaluation process was refined and started.

The user evaluation of speech-to-translated-text is covered in section 4.7 on MT components.

We briefly describe the speech-to-speech module here. It has not yet been integrated in the user-oriented testing platforms plain X and OSS, but users were able to do a first assessment of the voices through a specific LIA evaluation application.

The evaluation's modus operandi is to play back corresponding pairs of news segments – one in the original language and one in the target language. Testers are asked to assess to what degree the original voice's expressivity has been transferred to the target language.

Two assessments are envisaged, both using the Likert scale. First, the *degree* of expressivity from 1 ('The target language shows no expressivity') to 5 ('The target language shows a human-like expressivity'). Second, the *fidelity* of expressivity, which assesses how truthfully the original's expressivity was transferred to the target language, from 1 ('the target segment's expressivity does not match the original') to 5 ('the target segment's expressivity matches the original').

During the SELMA User Day in October 2022, Deutsche Welle users did such user testing to specifically assess the expressivity from one language into another.

In **Year 3**, a full user evaluation is foreseen, where we evaluate and compare the different processes. This means:

- the quality of the traditional ASR modules, from LIA, FhG and other available engines, as a basis for subsequent machine translation
- improvement output through ASR enhancement functions such as punctuation
- LIA's speech-to-translated text module compared to an ASR followed by MT
- LIA's speech-to-speech module compared to a traditional workflow of ASR + MT + synthetic voice

The evaluations will be done at different levels: a separate UI as provided by the developers, integration in plain X and in SELMA OSS. Users will record their findings in user questionnaires which will then be analyzed to provide comparative data.

4.7 TTS components

In **Year 1** we set the requirements for customizing synthetic voices for some of DW languages through a collaboration between DW and the University of Avignon, and selecting editorial departments and specific editors to be involved.

In **Year 2**, a total of eight voices of Brazilian DW journalists were cloned to develop the Brazilian text-to-speech component. This involved getting approval of the editors and collecting a dataset with audio and scripts with voices of the selected editors. These were collected by DW and handed over to LIA, who trained synthetic voices using that dataset.

On several iterations, the voices were assessed by the DW team and feedback was provided to the developers in terms of fluency, pronunciation accuracy and natural sound, including on interruptions in the output for certain voices, background noises (due to training from real content), robotic sounds, etc.. The assessment was repeated with new versions. The customized voices were integrated in the OSS, in plain X and in the podcasting application. Screenshots are included in the sections on the OSS and plain X platforms, see Figure 4.

In **Year** 3, we will do a full assessment of the enhanced synthetic voices by means of user questionnaires. This will allow us to do A/B listening comparisons with test users by playing back the same news segment, in one instance spoken by the journalist herself and in the other instance created through TTS. Test users can then rate each example on a Likert scale from 1 ('This sample sounds too robotic to be listened to') to 5 ('this sample sounds as natural as I would expect from a human reader'). Different aspects, such as pronunciation accuracy, naturalness, rhythm, flow, speech melody, intonation, pitch, etc, will be assessed.

4.8 User Scenario Evaluation

In **Year 1**, we defined the use cases and the user scenarios. D1.1 - Use Case Description and Requirements has identified 22 scenarios, which are part of our evaluation effort. This relates to functionality testing (at platform level) with a specific purpose, namely that of each of the scenarios. As stated in the Use Case Description, 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.

Evaluation is aimed at assessing and measuring their usability, accuracy and improvement over time.

Table 3 (User Scenarios) shows the current status in **Year 2,** listing the targeted scenarios, and provides a brief description for each and the focus of the evaluation. The last column indicates if evaluation of the particular scenario was done in the reporting period.

User Scenarios in Detail

#	Scenario ID	Scenario Description	Focus Evaluation	Evaluation Y2
1	Monitor Sources SEL-Sc-001	The user and language team specifies the input source(s) they wish to monitor through the system.	Functionality	Y
2	Ingest Media Item SEL-Sc-002	The system ingests media items from the sources.	Functionality	Y
3	Select Media Item SEL-Sc-003	The system selects media items and shows them to the user based on specific preferences.	Functionality	Y
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	Y
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	Y
7	Create Translation SEL-Sc-007	The system creates a translation for an individual AV media item.	Functionality, Accuracy, Gradual Improvement	Y

8	View Cluster/ Entity SEL-Sc-008	The user views the details of a cluster and/or an entity.	Functionality	Y
9	View Individual Media Item SEL-Sc-009	The user views an individual media item in relation to a cluster or entity.	Functionality	Y
10	Select Preferences SEL-Sc-0010	The user sets their preferences in the system.	Functionality, Usefulness	Y
11	Conduct Search SEL-Sc-0011	The user can search for an item in the system.	Functionality, Relevance of results	Y
12	Save Cluster / Individual Media Item SEL-Sc-0012	The user can save a cluster or an individual media item in the system where it is stored for more than a predefined set of time.	Functionality	Y
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	Y
14	Train System SEL-Sc-0014	The user can train the system in relation to the cluster generation.	Functionality, Accuracy, Gradual Improvement	Y
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	Y
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	Y
19	Generate Summary SEL-Sc-0019	The system generates a summary for each media item.	Functionality, Relevance and Accuracy, Usefulness, Gradual Improvement	Y
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	Y
21	Edit Transcription/Translation SEL-Sc-0021	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	Y
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 3** User Scenarios in Detail*

The goal for **Year 3** is to have all scenarios evaluated.

4.9 DW NLP Benchmarking

DW puts great effort in performing in-house benchmarking for the major NLP processes through direct use by the users, i.e., ASR (automated speech recognition) and MT (machine translation).

In **Year 1**, we established the benchmarking procedure and prepared the evaluation material.

This evaluation was started and is currently in process for all 32 DW languages and consists of both a human and an automated evaluation. A dataset was selected to serve as a baseline.

For ASR evaluation, we use up to three videos per target language, the videos are selected from the DW archiving system that already has an editorial script. This (manu)script is checked by an editor from the corresponding language department to assess accuracy. The automated evaluation is then performed by calculating the Word Error Rate (WER). We aim to use 3 videos for the benchmarking of each target language.

For the evaluation of MT, we use five videos that have been preselected in English and German, with the corresponding transcripts. We then request each editorial department to provide a reference text in the target language, sentence-aligned, for each of the videos, from either the English or German source transcript. Once we are provided with the reference texts and have obtained the output texts from each MT engine, we can do the actual human and automated assessment. The BLEU score was selected as a rating score.

The automated evaluation is supplemented by a user assessment, i.e., an evaluation of the quality of the transcription or machine translation output -- for all engines available to the DW team for the language (pair) being assessed -- by a native speaker proficient in the source language (as well as the target language in the case of MT). A rating is made for different aspects, including translation accuracy, punctuation and capitalization, fluency, completeness. Human evaluation is done by means of user questionnaires and Likert ratings of 1-5 on user satisfaction.

An initial partial dataset was created to cover few languages with which the first evaluations were done.

In **Year 2**, the reference dataset was expanded to more languages. Editorial departments were involved to provide the reference texts, i,e. (1) Check the accuracy of the transcript of the video selected in their target language against the audio content and (2) Provide a human translation for the English or German text of the five selected reference videos. Delivery of the edited content depends on the availability of the editors in the different language departments. Editors were asked to provide one ASR file and the first MT reference text as a priority, so the evaluation for their target language could be started. In this reporting year, editorial reference material was produced for Kiswahili, Serbian, Pashto, French, Spanish, Portuguese, Arabic, Indonesian, Urdu, Chinese, Bengali, Russian, Turkish, Ukrainian, Macedonian, Persian, Polish.

The list of NLP tools was expanded to include:

- 5 for ASR: Amberscript, Google, Azure, Speechmatics, and most recently Whisper
- 6 for MT: Google, Azure, Facebook, DeepL, eTranslation, GoURMET

We also asked the editors (native speakers) to assess the output of all available MT output and provide their feedback using the user questionnaire which was set up for this purpose, providing a satisfaction rating of 1 to 5. For some languages this meant evaluating up to 40 MT output texts, e.g. 5 texts in 4 tools (for instance Google, Azure, DeepL, Facebook) from two source languages (English and German).

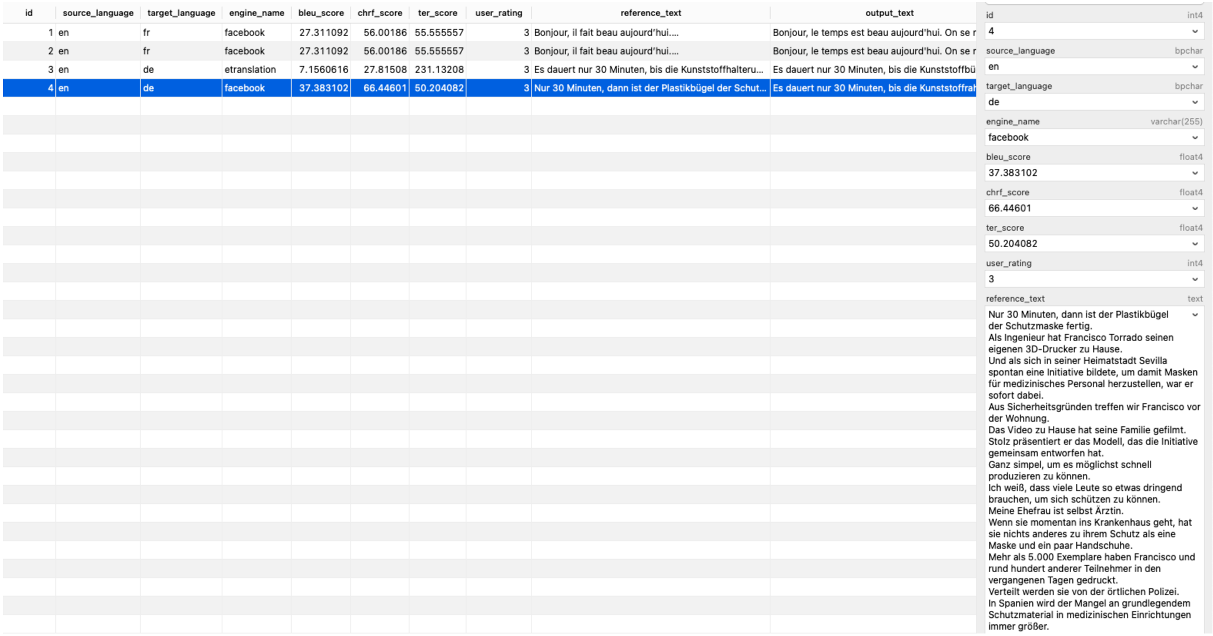
We also did back-to-back translations for all languages, so that we can do a comparative analysis even if no reference text is available (yet). This provided us with a basic evaluation of MT output of the different MT engines. It compares the English input and output text after an automated translation from English into the target language and then translating that output text back into English using the same provider, for instance Azure MT into English and translating that output back into English with Azure. This is also shown in Figure 1.

The automated evaluation process was refined, and it was decided to expand the MT automated ratings to three metrics for each language pair: BLEU, chrF and TER (Translation Error Rate). This provides a more reliable evaluation output.

To support the automatic evaluation of both ASR and MT tasks, the development of a web application was started and a first version is ready and in use. This tool allows users to upload both reference and output texts, calculate the metrics and store results into a database. The user can also obtain MT engine output from a source text directly and perform the automatic evaluation for MT if a reference text is available.

In addition to ASR and MT, we did some benchmarking on speech synthesis. This is explained in section 4.9.

In **Year 3**, the BM web application will be enhanced. A user will also be able to upload a video and obtain benchmarking results for ASR as well.



id	source_language	target_language	engine_name	bleu_score	chrf_score	ter_score	user_rating	reference_text	output_text
1	en	fr	facebook	27.311092	56.00186	55.555557	3	Bonjour, il fait beau aujourd'hui....	Bonjour, le temps est beau aujourd'hui. On se r
2	en	fr	facebook	27.311092	56.00186	55.555557	3	Bonjour, il fait beau aujourd'hui....	Bonjour, le temps est beau aujourd'hui. On se r
3	en	de	etranslation	7.1560616	27.81508	231.13208	3	Es dauert nur 30 Minuten, bis die Kunststoffhalteru...	Es dauert nur 30 Minuten, bis die Kunststoffbü
4	en	de	facebook	37.383102	66.44601	50.204082	3	Nur 30 Minuten, dann ist der Plastikbügel der Schut...	Es dauert nur 30 Minuten, bis die Kunststoffrah

id int4
4

source_language bpchar
en

target_language bpchar
de

engine_name varchar(255)
facebook

bleu_score float4
37.383102

chrf_score float4
66.44601

ter_score float4
50.204082

user_rating int4
3

reference_text text
Nur 30 Minuten, dann ist der Plastikbügel der Schutzmaske fertig.
Als Ingenieur hat Francisco Torrado seinen eigenen 3D-Drucker zu Hause.
Und als sich in seiner Heimatstadt Sevilla spontan eine Initiative bildete, um damit Masken für medizinisches Personal herzustellen, war er sofort dabei.
Aus Sicherheitsgründen treffen wir Francisco vor der Wohnung.
Das Video zu Hause hat seine Familie gefilmt.
Stolz präsentiert er das Modell, das die Initiative gemeinsam entworfen hat.
Ganz simpel, um es möglichst schnell produzieren zu können.
Ich weiß, dass viele Leute so etwas dringend brauchen, um sich schützen zu können.
Meine Ehefrau ist selbst Ärztin.
Wenn sie momentan ins Krankenhaus geht, hat sie nichts anderes zu ihrem Schutz als eine Maske und ein paar Handschuhe.
Mehr als 5.000 Exemplare haben Francisco und rund hundert anderer Teilnehmer in den vergangenen Tagen gedruckt.
Verteilt werden sie von der örtlichen Polizei.
In Spanien wird der Mangel an grundlegendem Schutzmaterial in medizinischen Einrichtungen immer größer.

***Figure 13** Example of database for MT evaluation*

The automated procedure for the automated assessment enables a very fast and consistent evaluation, which is essential in case of updates or newly available engines.

We will also explore ways of providing automated support for human evaluation of ASR and MT, with an analysis of the user input coming from the user questionnaires, making the process much faster and reliable.

We will further expand the reference dataset, involving more editors, native speakers of the target language, aiming at having a human translation of all five reference texts for MT for all DW languages and three reference texts in each of the 32 DW languages. This can be extended to other languages.

We also aim at getting a human evaluation for all these languages for MT and ASR output to supplement the automated evaluation outcome.

In the end, this benchmarking will produce a good overview of which engines provide the best results and are the most appropriate ones for use in a productive editorial environment. It compares major commercially available engines, open-source tools and those developed within the SELMA project. 32 languages are currently under assessment and have been at least partially evaluated. The automated evaluation process allows for an efficient updated assessment in case of new or enhanced tools.

5. Timeline

We are in line with the estimated timeline that was set in D5.1 - Evaluation Plan. Currently we are at M24.

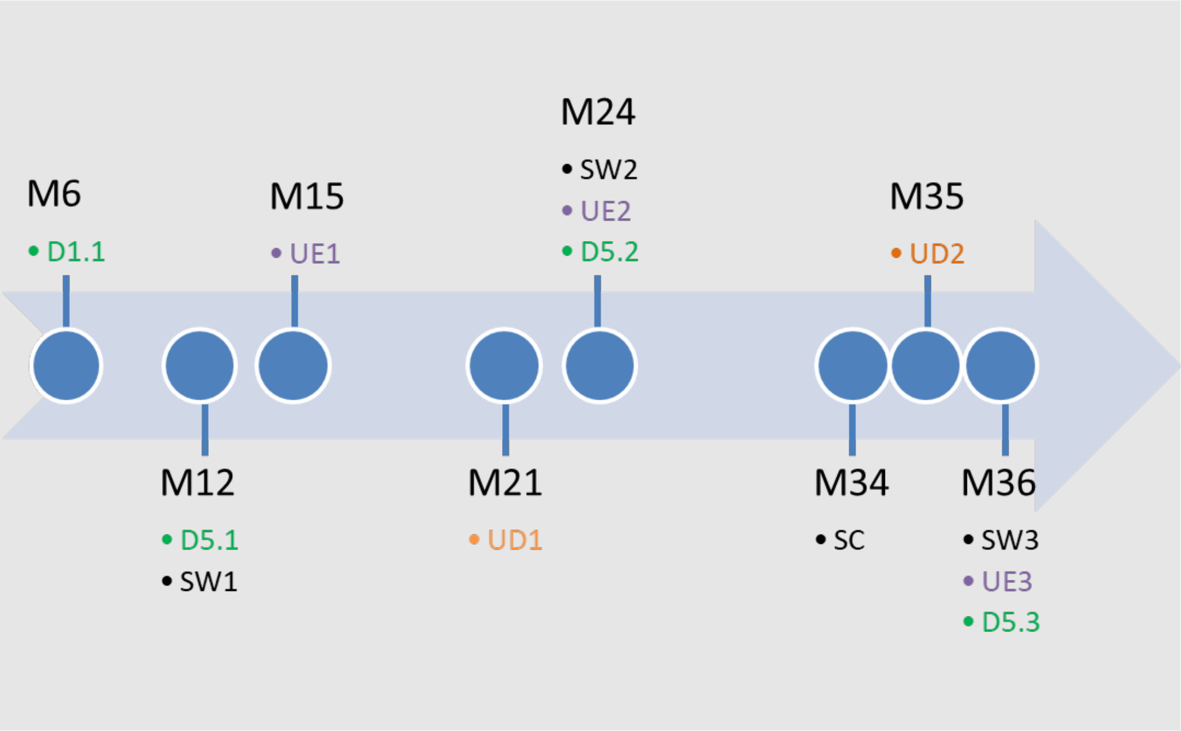


Figure 14 *Broad Timeline of Evaluation Activities Planned*

Timeline Legend:

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

Software releases and user evaluation ware done as planned. Our first SELMA User Day was held on 12 October 2022 in Bonn. More details are provided in this deliverable, section 2 and in D6.5 - Interim Impact Report.

6. Conclusion

This document follows up on D5.1 - Evaluation Plan and describes the evaluation activities in Y2 of the SELMA project. It combines contributions from all consortium partners and relates to other deliverables, including D1.1 - Use Case Descriptions and Requirements, D2.4 - Intermediate progress report on continuous massive stream learning, D2.5 - Intermediate release of stream learning and entity linking capabilities, D2.6 - Intermediate release of segmentation, summarization and news classification capabilities, as well as D3.4 - Intermediate progress report on speech and natural language processing, D3.5 - Intermediate release of transcription, punctuation and translation, voice synthesis capabilities, D3.6 - Intermediate release of post-editing and user feedback capabilities.

It provides an update of the Excel tracking sheet, our central evaluation planning and tracking tool used at consortium level to manage assessment by technology partners as well as user partners and at different levels.

In this phase, technical evaluation focused on ASR, speech to text, NER/NEL, summarization, integration of demonstrators and orchestration.

User assessment efforts targeted NER annotation, speech to text and speech translation, platform usage and usability for media monitoring and for subtitling, the podcasting and diversity use cases, in addition to benchmarking NLP tools in order to select the best available service for a large number of languages.

The final evaluation report will cover the ongoing testing described here and look at the final outcome in terms of KPI’s and TRL.

This D5.2 Interim Evaluation Report uses input from D5.1 - Evaluation Plan and will be followed by D5.3 - Final Evaluation Report.

7. Annex

7.1 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)
chrF	Character n-gram F-score (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)
TER	Translation Error Rate (measurement for MT)
ToC	Table of Contents
UCx	Use Case x
UD	User Day
UE	User Evaluation
UI	User Interface
UX	User Experience
WER	Word Error Rate (measurement for ASR)