Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D6.4 Interim Impact Report

Work Package	6
Responsible Partner	DW
Author(s)	Ksenia Skriptchenko
Contributors	Kay Macquarrie
Reviewer	Christoph Schmidt
Version	1.0
Contractual Date	31 December 2022
Delivery Date	22 December 2022
Dissemination Level	Public

Version History

Version	Date	Description
0.1	3/11/2022	Initial Table of Contents (ToC)
0.3	15/11/2022	Initial input
0.4	15/12/2022	Internal review
1.0	22/12/2022	Finalization and Submission

Executive Summary

The SELMA mission statement shows a central impact objective:

Shaping AI speech and text technologies for media & the newsroom

A major focus is to convince potential user partners about SELMA output and exploit components and platforms.

This deliverable, "D6.4 Interim Impact Report," outlines SELMA's plan for connecting with the relevant players in the commercial, academic, and general public sectors as well as for achieving the intended impact on its target audience.

This document has three sections:

- Dissemination Plan for SELMA: Outlines target audiences and defines a dissemination strategy based on the three phases of inform, involve, and convince.
- Communication Plan for SELMA: Provides a range of communication tools for a commanding visual presence and a unified project identity. It offers reporting measures as well as a basic sustainability strategy.
- Exploitation Strategy for SELMA: Examines the outcomes of the first applications of the findings and the exploitation of SELMA NLP components, including an open-source platform.

SELMA is a “Research and Innovation Action” project. Therefore, the project has two impact focuses:

a) advancing the state-of-the-art in various NLP related tasks and technologies through improvements in research, and

b) bringing technology to the market through tools, components and an open-source platform.

The Dissemination and Communication Plan will be updated within the course of the project development and will be finalized at the end of the project (M36).

Table of Contents

Table of Figures

Table of Tables

1. Introduction

This Impact Report is divided into three main parts: Dissemination, Communication, and Exploitation.

Section 1 (Dissemination) outlines plans and shows efforts to inform, inspire and involve future SELMA users and the research community. SELMA's dissemination goal is to engage potential users, HLT (Human Language Technology) providers, and early adopters to establish a feedback culture and build a network with other projects, researchers, and technology users.

The communication strategy is laid out in **Section 2 (Communication)**, which also identifies the primary channels and activities for communication. Additionally, it highlights the initiatives already made to increase public knowledge of the project and its advancement.

A summary of exploitation activities is provided in **Section 3 (Exploitation)**.

The dissemination and communication plan as well as the exploitation strategy will be constantly revised, updated, and refined throughout the course of the project. There will be one follow-up deliverable (D6.6 Final Impact Report in M36).

2. Dissemination

The purpose of this section is to provide a project dissemination strategy by highlighting targeted groups and communities, defining internal dissemination/communication guidelines and procedures, outlining the foreseen channels, and reporting on the efforts in the first project year.

This section describes the efforts of defining, identifying, and reaching our target audience, focusing on its two primary target groups, i.e., the scientific communities and the media world.

It provides the plan of selecting, setting up and supplying the right dissemination channels and a survey of the dissemination events to promote the results in the related fields of research.

Following the dissemination activities framework outlined in the DoA, SELMA adopts a multi-channel and multi-target approach. SELMA pursues a clearly defined strategy, which will be outlined and specified further below in the sub-chapters.

2.1 Target Groups

Overall, we divide our target audience for dissemination activities as follows, and will address primarily:

- Broadcasting and media world
- Industries using language processing technologies
- Translation agencies
- Media monitoring organizations
- Industries in need of monitoring multilingual content across the world media
- Scientific and research community
- Stakeholders and their networks
- Policy makers and interest groups
- Human Language Technology users
- General public (interested users)

2.1.1 Network

All of the SELMA consortium's partners have established networks and exert every effort to inform and engage with their intended markets both on an individual basis and collectively.
The SELMA consortium also intends to support the Big Data Value Association in any actions related to the project's activities.

2.1.2 User Group

The User and Advisory Board serves as an advisor to the Innovation Manager and the project's Steering Board.

In March, 24th 2022 SELMA had its first User Group meeting. It was organized as an interactive online meeting and brought together NLP and media experts from many organizations including BBC, University of Edinburgh, COFINA, Visapress and AICEP. The status of SELMA developments and Use Cases were showcased and feedback from the user group was discussed and collected. As an outcome the Advisory Board was formed, which is a smaller subset of the User Group and consists of representatives of these organizations: BBC, EBU and University of Edinburgh.

User group activities include identifying the strong and weak points with respect to the objectives of the project (with emphasis on the innovation objectives) and providing recommendations. Furthermore, the members of the User Group help us maximize our industry outreach, serving as links between the consortium and external key industry players. Currently, around 15 representatives from research organizations, European media companies and technology providers are forming the SELMA user group.

2.2 Strategy

The plan identifies the main initiatives and channels that will support the project goals. Events, the creation of materials to describe and demonstrate project accomplishments, and targeted publications are all governed by the dissemination strategy. Three significant stages have been established by the project consortium.

	Y1	**Y2**	Y3
Phase 1	Inform & inspire		
Phase 2		**Involve & contribute**	
Phase 3			Share & Convince

Table 1 *Three dissemination stages and progress at the end of the second year*

2.2.1 Inform and inspire

The ***"inform and inspire"*** phase already started in the first year and will be active during the whole lifetime of the project. The focus in the first phase is on:

- Outlining the project's vision, aims and goals
- Setting up the dissemination channels and activities to spread the word and to inform target audiences about SELMA and its major objectives
- Introducing and interacting with relevant communities, which might differ in the scope of different use cases

2.2.2 Involve and contribute

The ***"involve and contribute"*** part started in the second year and will continue until the end of the project. This phase will be consisting of the following steps:

- Identifying key influencers to involve into the feedback loop, in testing early prototypes, sharing research results
- Utilizing early adopters as multipliers and to spread further awareness
- Obtaining user feedback on development and creating solutions for obstructions

2.2.3 Share and convince

The last part of the dissemination strategy ***"share and convince"*** starts in the third year and mainly focuses on:

- Demonstrating progress by making available a variety of open-source resources and project outcomes
- Showing prototypes and innovative features to targeted audiences and third parties
- Engaging with target groups and individual users to support the SELMA exploitation activities

2.3 Dissemination

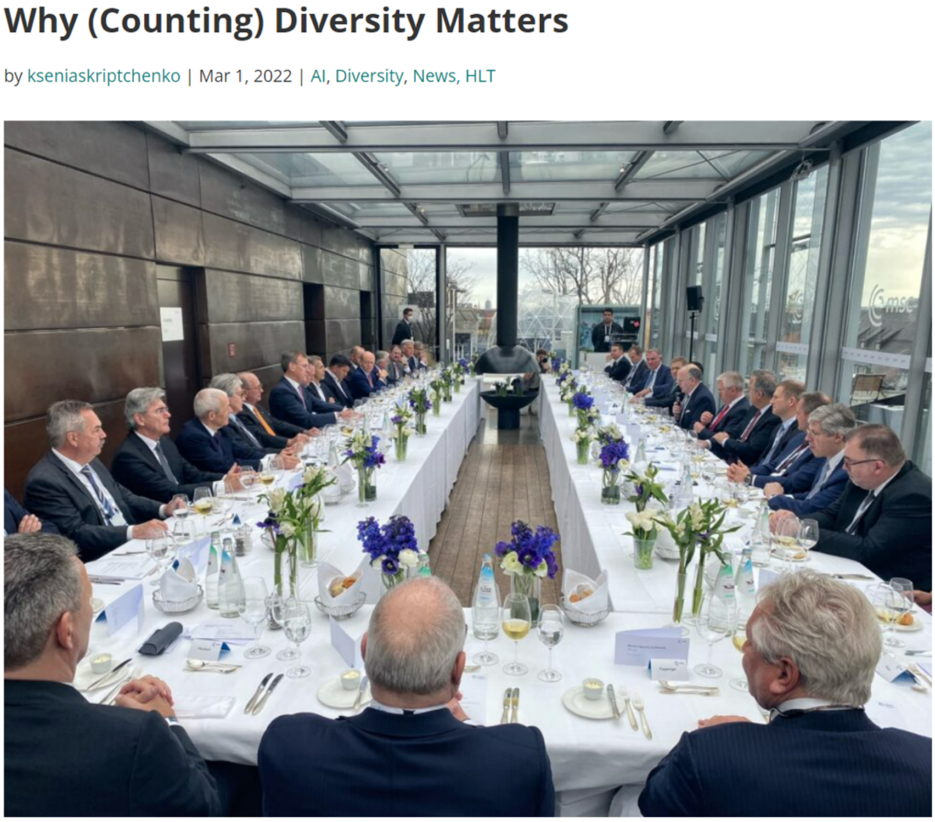
The general SELMA dissemination strategy and approach were discussed in the chapter above. The next part provides more information on how we intend to accomplish the predetermined objectives as well as what has been planned and accomplished throughout the project's second year.

2.3.1 Channels (Website, Social Media)

Website and Blog

An important channel of the SELMA communication and dissemination strategy is its website. The SELMA homepage can be reached at *www.selma-project.eu.* It provides information on the project, main goals and project partner descriptions, and contact person information.

The blog feature of the homepage serves as the SELMA repository of the pertinent contributions to relevant artificial intelligence and human language technologies made by consortium partners. The public at large is the primary audience.



Why (Counting) Diversity Matters

by kseniaskriptchenko | Mar 1, 2022 | AI, Diversity, News, HLT

Figure 1 *The most viewed SELMA blog post in 2022 with 480 post views*

The following content was published in the second year of the project:

- Priberam contribution on Trustworthy AI
  https://selma-project.eu/2022/12/19/on-the-path-to-the-responsible-ai/

- Fraunhofer contribution on Speaker Diarization
  https://selma-project.eu/2022/12/13/who-spoke-when/
- IMCS UL contribution on TokenQueue Service
  https://selma-project.eu/2022/06/30/need-computing-resources-take-a-queue-token/
- LIA contribution on Self-Supervised Learning
  https://selma-project.eu/2022/05/26/how-to-satisfy-data-hungry-machine-learning/
- DW contribution on Importance of Diversity in Media
  https://selma-project.eu/2022/03/01/why-counting-diversity-matters/

It is constantly encouraged for all SELMA partners to actively contribute content to the website, especially for the blog section.

In the final year we will offer more frequent updates, such as:

- various documentations on project activities and outcome,
- links to prototypes, demos and tutorials,
- code, datasets and resources.

Social Media

This section lists social media-based dissemination approaches and describes corresponding strategies.

YouTube

This platform was chosen to present our findings in the form of videos. Furthermore, we use our YouTube presence to reach out to new contacts and target new collaborations. The first YouTube contribution was an image video that gave our target audience a brief but in-depth overview of the project.

SELMA - an EU research project on language technologies

Figure 2 *SELMA's image video as seen on SELMA's YouTube channel*

This image video was made available in English. Part of the video was artificially dubbed using SELMA's own plain X application. Subtitles in English, German, and Russian were also created using plain X. The following table summarizes SELMA's video output in the second project year and gives a brief overview of video topics.

Videos	**Description**	**URL**
SELMA Image Video	SELMA helps media monitors and journalists make sense of huge content streams, also making audiovisual output more accessible through transcription, translation, voice over and subtitling.	URL
New Podcast Creation (Demonstration)	One concrete application of the SELMA project is the Podcast Creator. Its purpose is to create a podcast almost on-the-fly using search, summarization and speech synthesis techniques with customized SELMA voices.	URL
"The Footprint of AI & NLP Technology in the Media - What Comes Next?" (User Day Panel Discussion)	Stephanie Bradford (DW Diversity), Kirsten Radtke (DW Informations), Guntis Bārzdiņš (University of Latvia), Afonso Mendes (Priberam), moderated by Olga Kisselmann (DW Research & Cooperation Projects), discussing the topic during the User Day.	URL

"Data Representation Matters - Counting Diversity" (User Day Panel Discussion)	Mirjam Gehrke, senior editor from Deutsche Welle highlights the needs of a broadcaster to meet strategic goals in this field. Kay Macquarrie, coordinator of the SELMA project, shows how SELMA technology can help to count diversity numbers based on NLP and data analysis.	URL
"plain X - The Four in One NLP Tool" (User Day Demonstration)	plainX in a nutshell: plainX is an integrated platform combining a task-based workflow with access to powerful HLT (human language technologies).	URL
"Monitio - A Multilingual Media Monitoring Tool (User Day Demonstration)	Monitio searches and analyses content (currently up to 200,000 articles a day from around the world) to deliver information and indications for better decision-making.	URL
"Artificial Intelligence (HLT) & The Newsroom" (User Day Panel Discussion)	Bird's eye perspective from Peggy van der Kreeft on how AI at Deutsche Welle evolved (featuring various EU projects), basically covering the past decade. Includes an outlook into the future.	URL

Table 2 *SELMA related videos, produced and posted in the 2nd year*

GitHub

We have made some of the SELMA software and applications available on GitHub. It is therefore possible to use and try out the SELMA open-source platform for NLP tasks including transcription, translation and voice-over in many languages through this channel. The following is a link to the browser-based platform: https://selma-project.github.io/

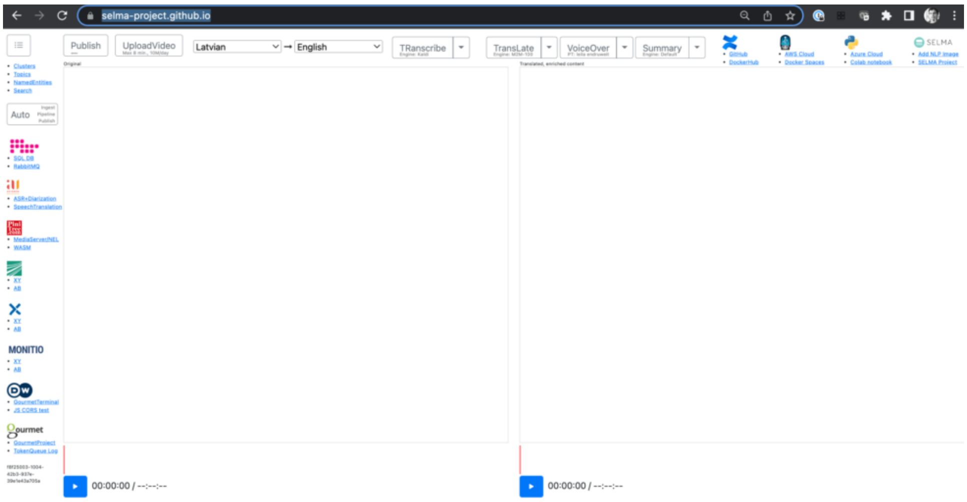


Figure 3 *SELMA's open-source platform*

The open-source applications that have also been made available are listed in the table below.

Module	**Description**
ml4audio	Audio, NLP, ML with huggingface, nvidia/nemo, speechbrain
SELMA-project.github.io	UC0 Frontend
plotly-dash-subtitles-creator	Semi-automated workflow to create subtitles for video via plotly dash

Table 3 *GitHub Repository*

The following table gives a brief report on the status of planned dissemination tools and materials, explained and described above.

Channels	**Status**	**Year**
Website	set-up/ongoing updates	Y1/Y1-3
Blog	set-up/ongoing updates	Y1/Y1-3
Twitter	set-up/ongoing updates	Y1/Y1-3
LinkedIn	set-up/ongoing updates	Y1/Y1-3

Github	set-up/ongoing updates	Y1/Y1-3
YouTube	set-up/ongoing updates	Y2/Y2-3

Table 4 *Dissemination Channels*

Dissemination materials will be updated during the project if required.

2.3.2 Events

An immensely important part of the dissemination strategy is being present at relevant events for the SELMA project work. It helps us to stay informed and up to date in the scientific areas, present project achievements and results, meet relevant stakeholders for future collaboration and cooperation, and prepare for exploitation activities. Given the pandemic restrictions, we have been attending industrial and academic events, the majority of which were held digitally.

A list of attended or organized conferences (academic & industry), workshops, and other dissemination events in the first year of the SELMA project is provided in the table below. User Days / Events are listed separately.

#	Date	Title	Partner
1	15.02.2022	Hipeac Webinar	DW
2	21.-23.03.2022	Propor 2022	Priberam
3	22.3.2022	Online Workshop of the Interinstitutional Task Force on Speech Recognition (European Union)	LIA
4	22.-27.05.2022	IEEE ICASSP 2022	LIA
5	22.-27.05.2022	ACL 2022 + IWSLT 2022	LIA
6	10.06.2022	Fifth Workshop on Narrative Extraction From Texts held in conjunction with the 44th European Conference on Information Retrieval (ECIR 2022)	Priberam
7	13-17.6.2022	JEP 2022 Journées d'étude sur la parole	LIA
8	15.06.2022	27th International Conference on Natural Language & Information Systems (NLDB 2022)	IMCS
9	17.06.2022	EBU Workshop (EBU Access Services Expert Event)	DW

10	20-25.06.2022	LREC 2022 - International Conference on Language Resources and Evaluation	Fraunhofer
11	27.6-1.7.2022	TALN 2022 Conférence sur le Traitement Automatique des Langues Naturelles	LIA
12	27.6-5.8.2022	JSALT Workshop 2022	LIA
13	18-22.9.2022	ISCA Interspeech 2022	LIA
14	14.10.2022	Fête de la Science	LIA
15	14.10.2022	Presentation of SELMA DockerSpaces technology to AI-Lighthouse (HORIZON-CL4-2022-HUMAN-02-02) consortium meeting at Oulu University	IMCS
16	25.11.2022	Diversity Use Case presentation for ARD AI working group	DW
17	07.-12.12.2022	EMNLP 2022 (The 2022 Conference on Empirical Methods in Natural Language Processing)	Priberam
18	14.12.2022	Second BDVA workshop with ICT-51 projects	DW
19	15.12.2022	LOCDOC Master Class on plain X	DW, Priberam

Table 5 *List of Dissemination Events Year 2*

As shown in the table above, DW participated in the BDVA workshop and actively contributed to the collaboration of EU projects in the field of Big Data.

2.3.3 User Events (User Board & User Day)

For SELMA products and results to stay visible and accessible for our stakeholders in different target groups, we regularly organize user events.

As important as the attendance of other technology and research-related events is, the hosting of SELMA-related events gives us extensive opportunities to focus on the dissemination of the results of the SELMA project. By presenting demonstrators and by meeting face-to-face relevant stakeholders for future collaboration and cooperation, we are able to show the relevance of SELMA project achievements for the HLT community.

User & Advisory Board Meeting

On March 24th, 2022, SELMA had its first User Group meeting. It was organized as an interactive online meeting and brought together 31 NLP and media experts from many organizations, including BBC, the University of Edinburgh, COFINA, and AICEP.

Figure 4 *First SELMA User Group Meeting announcement*

The status of SELMA developments and Use Cases were showcased and feedback from the user group was discussed and collected. As an outcome, the Advisory Board was formed, which is a smaller subset of the User Group and consists of representatives of these organizations: BBC, EBU and University of Edinburgh.

First SELMA User Day

The first SELMA User Day had the maxim to "Simplify Life in the Newsroom" with the use of AI and Language Technology. The hybrid event (which also featured a live stream and remote attendants) brought together professionals from all sorts of European media organizations, including ARTE, the BBC, Lusa, and Priberam. It took place on October 13, 2022, in Bonn and was hosted by Deutsche Welle.

***Figure 5** First SELMA User Day in Bonn, Germany*

The User Day had three main parts: interactive presentation sessions, a panel discussion and a workshop/ poster presentation area. The morning block consisting of interactive presentations gave insights into the latest advances into SELMA-related multilingual language technology, and how it's applied in the field of journalism and media production. At noon a panel discussion on the "Footprint of AI & NLP Technology in the Media" followed. In the afternoon, participants were invited to test various demo versions, including the two main Use Cases plain X and Monitio, the Podcast Creator, the SELMA open-source tool and a demo on how to segment audio files. LIA invited participants to provide direct feedback on the latest developments on speech synthesis of specific voices. Valuable feedback was being collected from the meeting. One of the participants stated concerning the Monition Diversity Use Case: "This gives me an instant overview of our [editorial] output – and our weak spots. A lot to think about!" In the course of the afternoon, LIA was able to collect enough participant feedback to further improve work on the speech synthesis.

During the User Day, SELMA could reach out to around 86 persons (up to 58 persons in the stream simultaneously and 28 people physically in the room). The presentations and the panel discussion are also available on demand and have received 182 views so far.

The second SELMA User Day is scheduled for November 2023. It will most likely take place in Madrid or Lisbon. The focus will be on targeting media people and commercial media companies.

2.3.4 Publications

Publications in the scientific and academic spheres are an essential component of SELMA dissemination activities. A list of 17 publications and research papers published in the second year is provided below.

#	Consortium Partner/ Names	Title	Published in	Open Access & Golden Standard (+URL)
1	*IMCS:* *Paulis Barzdins, Audris Kalnins, Edgars Celms, Janis Barzdins, Arturs Sprogis, Mikus Grasmanis, Sergejs Rikacovs, Guntis Barzdins*	***Metamodel Specialisation based Tool Extension***	Baltic Journal of Modern Computing	*Yes*
2	*LIA:* *Salima Mdhaffar, Valentin Pelloin, Antoine Caubrière, Gaëlle Laperrière, Sahar Ghannay, Bassam Jabaian, Nathalie Camelin, Yannick Estève*	***Impact Analysis of the Use of Speech and Language Models Pretrained by Self-Supersivion for Spoken Language Understanding***	Proceedings of the 13rd Language Resources and Evaluation Conference (LREC)	*Yes*
3	*LIA:* *Gaëlle Laperrière, Valentin Pelloin, Antoine Caubrière, Salima Mdhaffar, Sahar Ghannay, Bassam Jabaian, Nathalie Camelin, Yannick Estève*	***The Spoken Language Understanding Media Benchmark Dataset in the Era of Deep Learning: data updates, training and evaluation tools***	Proceedings of the 13rd Language Resources and Evaluation Conference (LREC)	*Yes*
4	*LIA:* *Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, Yannick Estève*	***Speech Resources in the Tamasheq Language***	Proceedings of the 13rd Language Resources and Evaluation Conference (LREC)	*Yes*
5	*LIA:* *Gaëlle Laperrière, Valentin Pelloin, Antoine Caubrière, Salima Mdhaffar, Nathalie Camelin, Sahar Ghannay, Bassam Jabaian, Yannick Estève*	***Le benchmark MEDIA revisité : données, outils et évaluation dans un contexte d'apprentissage profond***	Journées d'Études sur la Parole - JEP2022	*Yes*
6	*LIA:* *Hang Le, Sina Alisamir, Marco Dinarelli, Fabien Ringeval, Solène Evain, Ha Nguyen, Marcely Zanon Boito, Salima Mdhaffar, Ziyi Tong, Natalia Tomashenko, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet,*	***LeBenchmark, un référentiel d'évaluation pour le français oral***	Journées d'Études sur la Parole - JEP2022	*Yes*

	Solange Rossato, Didier Schwab and Laurent Besacier			
7	*LIA:* *Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab and Laurent Besacier*	***Modèles neuronaux pré-appris par auto-supervision sur des enregistrements de parole en français***	Journées d'Études sur la Parole - JEP2022	Yes
8	*LIA:* *Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, Yannick Estève*	***ON-TRAC Consortium Systems for the IWSLT 2022 Dialect and Low-resource Speech Translation Tasks***	Proceedings of the IWSLT 2022 (ACL Anthology)	Yes
9	*LIA:* *Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, Shinji Watanabe*	***Findings of the IWSLT 2022 Evaluation Campaign***	Proceedings of the IWSLT 2022 (ACL Anthology)	Yes
10	*Priberam:* *João Santos, Afonso Mendes, Sebastião Miranda*	***Simplifying Multilingual News Clustering Through Projection from a Shared Space***	Proceedings of the Text2Story (Fifth Workshop on Narrative Extraction from Texts held in conjunction with the 44th ECIR, 2022)	Yes
11	*LIA:* *Marcely Zanon Boito, Laurent Besacier, Natalia Tomashenko, Yannick Estève*	***A Study of Gender Impact in Self-supervised Models for Speech-to-Text Systems***	Proceedings of the INTERSPEECH 2022	Yes

12	*LIA:* *Salima Mdhaffar, Jarod Duret, Titouan Parcollet, Yannick Estève*	***End-to-end model for named entity recognition from speech without paired training data***	Proceedings of the INTERSPEECH 2022	Yes
13	*Priberam:* *Pedro Henrique Martins, Zita Marinho, Andre Martins*	***∞-former: Infinite Memory Transformer***	Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)	Yes
14	*LIA:* *Gaëlle Laperrière, Valentin Pelloin, Mickaël Rouvier, Themos Stafylakis, Yannick Estève*	***On the Use of Semantically Aligned Speech Representations for Spoken Language Understanding***	Publication in the SLT 2022: IEEE Workshop on Speech and Language Technology 2022	Yes
15	*IMCS:* *Eduards Mukans, Gus Strazds, Guntis Barzdins*	***RIGA at SemEval-2022 Task 1: Scaling Recurrent Neural Networks for CODWOE Dictionary Modeling***	Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), July 14-15, 2022. Seattle, United States	Yes
16	*IMCS:* *Arturs Znotins, Roberts Dargis, Normunds Gruzitis, Guntis Barzdins, Didzis Gosko*	***RUTA: MED - Dual Workflow Medical Speech Transcription Pipeline and Editor***	Proceedings of the NLDB 2022: Natural Language Processing and Information Systems. Lecture Notes in Computer Science book series (LNCS, volume 13286)	Yes
17	*Priberam:* *Diogo Pernes, Afonso Mendes, André F.T. Martins*	***Improving abstractive summarization with energy-based re-ranking***	Proceedings of the GEM workshop at EMNLP 2022	Yes

Table 6 *List of Publications Year 2*

2.3.5 Innovation Radar

Three SELMA project outcomes (as listed in the table below) were submitted and selected by the EU Innovation Radar Platform. In early 2023, the applications will be added to a list of EU-based innovations.

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

***Table** 7 List of Innovations submitted to the EU Innovation Radar*

3.Communication

SELMA's communication initiatives are described in this section. There are three sections to it. The communication strategy establishes the project identity and stipulates rules for internal and external communication.

This section's "Evaluation" subsection gauges and assesses the project's second-year performance. Additionally, it makes recommendations for when and where strategy and activities should be modified in light of the communication success metrics table.
Last but not least, the "Sustainability" chapter offers details on how the project's outcomes and results can be accessed and made available via publically accessible channels beyond its active phase.

3.1 Strategy

SELMA's communication strategy describes the channels and methods used to efficiently implement communication:

- within the project (internally),
- with peer researchers, related projects,
- with end users and the public at large.

It is important to ensure that the entire consortium is informed of the development status and achievements, also of the challenges to overcome, and the efforts invested. Equally important is communication towards the relevant research community, related projects, and our main target group consisting of media and broadcasting professionals to inform and ensure that the SELMA results can be used in professional communities outside of the consortium and relevant feedback can be provided during the project's lifespan.

The SELMA output can be used by a wide range of users interested in consuming content in other languages, using subtitles, full-text transcripts, or voice-over applications. To ensure the continued use of the outcomes and results, it is success-critical to inform the possible end users and general public about advantages, limitations, and overall progress of the SELMA's outcome.

3.2 Evaluation & Reporting

The table below evaluates the progress on the communication means and outreach:

Activity	Planned (in total)	Description	Achieved in total (by M24)
Website Visits	24.000	Y1: 7.000 Y2: 7.000 Y3: 10.000	13,475
Twitter Followers	180	Y1: 60 Y2: 60 Y3: 60	85
Tweets	150	Y1: 50 Y2: 50 Y3: 50	89
LinkedIn Followers	90	Y1: 30 Y2: 30 Y3: 30	81
LinkedIn Posts	50	Y1: 15 Y2: 15 Y3: 20	29
Videos	10	Y1: 0 Y2: 5 Y3: 5	7
Poster	1	(updated version if required)	1
Flyer	1	(updated version if required)	1
Roll-Up	1	(updated version if required)	1
Events (participated and organized)	55	Y1: 15 Y2: 20 Y3: 20	39
Publications	11	Y1: 3 Y2: 4 Y3: 4	21
User Events	4	Y1: 0 Y2: 2 Y3: 2	2
GitHub Watches	30	Y1: 0 Y2: 0 Y3: 30	0

***Table 8** Overview Communications Means & Progress*

As shown in the table above, almost all previously set targets were met or even exceeded. With more than 20 publications in the second year we have significantly more scientific output than anticipated. The numbers on Twitter, which was previously seen as one of our primary communication channel, fell short of expectations. This is due to the current situation with Twitter's takeover. We are closely monitoring developments in order to respond quickly. In the future, we will place a greater emphasis on our LinkedIn channel and shift our communication there.

3.3 Sustainability

To ensure the availability of the work performed by the SELMA project, the consortium aims at keeping dissemination channels available for a period of at least three years after the end of the project. This will include the website, and the social media channels.

Although the editorial input will be kept to a minimum after the project, all channels of distribution shall be made and kept accessible when the project is completed.

4. Exploitation

This section of the deliverable gathers the information required and explores options to ensure the outputs of the SELMA can be exploited by consortium partners and others. As the exploitation roadmap, this document sets out the activities required for the successful exploitation of SELMA results. The exploitation activities started in the second year of the SELMA project and will be a focus activity in the third year, including the required IPR strategies.

4.1 Exploitation Ways

There are four main ways in which SELMA will be exploited:

- **Orchestration platform**: The focus is to provide an open-source big-data platform able to ingest and orchestrate the pipeline graph of NLP modules and apply stream learning and user feedback. During the project the partners will investigate further options beyond open sourcing it and consider revenue potential of the full platform or the most promising parts.
- **Component-based/individual system modules:** The SELMA components/ modules have a high potential value as improvements for existing services, or as the basis for other new services.
- **Integration into other projects/products:** Results will be also integrated in MONITIO project (an H2020 FTI project for AI powered Media Monitoring lead by Priberam) and will be a part of plain X service, a novel human language technology (HLT) platform developed by DW Innovation in cooperation with Priberam.
- **Knowledge**: Through continuous research, each consortium partner is learning and developing new techniques that can be applied to succeeding projects. It also can be utilized for designing new approaches and new services.

4.1.1 Exploitation: Open-source & commercial approach

SELMA is working on various output formats including software components, platforms and training data. A first overview of the exploitation approach (open-source versus proprietary) is shown in the following table.

#	Components	Open-source	Proprietary	Comment
1	ASR	y		
2	MT	y		
3	News Summarization	y		The EBR model is OS
4	NER, NEL, Discovery		y	
5	Automatic Post Editing			Not defined yet
6	News Clustering	y		
7	Topic Detection		y	
8	Speech Synthesis	y		
9	Alert System		y	
10	Indexation		y	
11	Search & Visualization (UI)		y	
12	Story Segmentation	y		
13	SELMA OSS platform (UC0)	y		
14	Monitio platform (UC1)		y	
15	plain X platform (UC2)		y	
16	NER Dataset	y	y	Partly open-source, partly proprietary due to content privacy

***Table 9** Exploitation overview of SELMA output*

4.2 Exploitation Plan

The initial exploitation of the consortium is shown in the following table.

Partner	Exploitation Plan
LIA	Avignon Université (LIA) will use the technologies developed in the project to improve both the university's expertise in enriched transcription and translation and will expand this to its research network. LIA has many collaborations with industrial partners that could be interested to the results of the SELMA project: for instance Vecsys/Bertin, a French company that develops vocal technologies on speech processing commercialized all over the world; Orange has continuously shown interest in semantic information extraction from speech, Airbus is also a regular LIA partner interested in speech and language processing, or the INA that needs automatic speech and language processing to manage millions of video documents. LIA researchers are largely involved in a Master's degree specialized on machine learning and language processing proposed by the Avignon Université. Research in the SELMA project will feed courses and (lab) seminars. Lastly, LIA aims to publish high-level scientific publications in major peer-reviewed international journals and conferences. LIA has also very strong relations to the Language and Speech Technology team of LIUM, a public research laboratory hosted in Le Mans Université. Some future collaborations between LIA and LIUM researchers will have a positive impact on the SELMA project.
DW	Deutsche Welle will be directly involved with the system as user partner. Its data is ingested into and processed through the platform. A high level of customization is envisaged, for instance, voice synthesis is trained on and customized for some DW news readers. Punctuation as well as summarization is trained on and tested with DW content. Deutsche Welle journalistic staff and the development team are involved from an early stage, to enable and facilitate early implementation and beta testing in-house. The SELMA system is compatible with an HLT platform currently under implementation at DW, resulting from the SUMMA and news.bridge projects. SELMA will ensure continuous improvement and sustainability of HLT applications in a large number of languages. Thus, Deutsche Welle intends to use the platform to process and distribute its content in more languages, serving also smaller language departments, and introduce automated subtitling and voice-over, with post-editing, as standard procedures. It will work with other broadcasters in its network for wider use and feedback.
Fraunhofer	Fraunhofer will use the technologies developed in the project in industry projects for the German media industry, especially for German public broadcasters. In its long-term collaboration with the public broadcaster WDR as the lead buyer of AI technologies for the ARD group, Fraunhofer will introduce the developed technologies to fully search the vast media archives of the public broadcasters. The technologies and advances described in the proposal are closely correlated to demands in the media industry, especially related to improved speech recognition, punctuation prediction, live subtitling and accurate speaker recognition. The results obtained in the project are consequently directly supporting public broadcasters in Germany to improve their programm, e.g. by more efficient search of their archives or faster production of media assets and subtitles, thus reducing costs and delivering news faster and more reliably. Besides the media industry, Fraunhofer is supporting German state parliaments to automatically subtitle their parliamentary debates to ensure barrier-free access for the Deaf and hard-of-hearing. Improvements in speech recognition quality and punctuation will directly benefit this and enable us to scale these subtitling efforts to more regional or even the federal parliament. Fraunhofer will exploit the transfer learning techniques developed in this project to scale to more languages as well as different branches of industry (e.g. banking, call center, health industry), a scaling which was previously not possible due to lack of sufficient industry specific training data.
IMCS	IMCS as the integrator of the SELMA project will continue to develop the technologies conceived in the project and will disseminate by making them commercially available for the broader user group. IMCS will also exploit project results by implementing them in the LETA news agency multilingual (Latvian, English, Russian media) video news production workflows and for joint knowledge graph and large language model technology development with the PiniTree.com startup. IMCS will also use the SELMA voice-over system to extend the media monitoring and storyline summarization system built in SUMMA to automate the generation of video from the news items and synthetic narration of the summary.

Priberam	Priberam is currently productizing the results of SUMMA in MONITIO and news.bridge in plain X and as such plans to incorporate the results of SELMA into those new offerings. Priberam will also incorporate results from SELMA in its line of NLP SaaS products, which are sold directly and through partners in Portugal, Spain and Brazil to clients which include the biggest media producers. Together with IMCS, Priberam intends to support the platform and its derived products giving commercial support to the open-source parts and reaching agreements with the other parties on the commercial use of specific components. This renewed product line and its case studies will be presented in international roadshows as well as conferences in areas covered by the project. Priberam will disseminate the results across media groups and media monitoring target user companies in European countries, reaching also organizations from Latin America and Africa.

Table 10 *Initial Exploitation Plans -still valid in Year 2*

4.3 Expected Technology Impacts

SELMA's output will include a platform with improved research and tools for the various NLP technologies. The general OSS platform and some components will be released as open source. The following table sums up the expected technology /components achievements as listed in the initial DoA.

Technology / Component	TRL (2020)	Expected TRL (2022)	Status M12	Status M24
Punctuation Recovery	5	7	as planned	as planned
Speaker Diarization	6	8	as planned	as planned
Speaker Recognition	6	8	as planned	as planned
Rich Automatic Speech Recognition (including named entities)	6	8*	as planned	as planned
Text Machine Translation	7	9*	as planned	external integration
Expressive and Personalized Voice Synthesis	5	7	as planned	as planned
Speech Machine Translation	5	7	as planned	as planned
Topic Labeling (from crosslingual transfer)	4	~~6~~, 8**	as planned	better than planned
Named Entity Recognition and Linking	6	8	as planned	as planned
Abstractive Summarization	3	6	as planned	as planned
Integration Platform (NLP Components and UX)	6	8	as planned	as planned

Integration Platform (Learning/Training of NLP and Automatic Redeployment)	3	7	as planned	as planned
*depending on the target languages and language pairs				
** TRL raised from 6 to 8 in Y2 due to significant progress				

Table 11 *Expected Technology / Components Improvements Overview*

4.4 IPR Management

For a clear IPR Management, we are setting up an overview table, in which each component is described using a set of descriptors:

- Component name
- Inputs from
- Outputs to
- Component lead partner
- Component contributors
- Brief description
- What it does (more detail)
- How it works (more detail)
- Key innovative aspects
- Potential applications
- Software & IPR status
- Terms & conditions of use
- Performance requirements
- Further documentation
- Alternatives
- Key contact(s)
- Potential applications of any components aside from SELMA
- The names and details of any libraries used within components and their IPR status

- Answering key questions developers wishing to exploit the component are likely to have, such as: How many users can be supported? What specific machines are required to run the code?
- The name of and links to alternative open-source components if a component will not be part of an open-source release

This work will start in the first half of 2023.

5. Conclusion & Outlook

In the second year of the project, we continued to show presence at various conferences and events (19 in total). With 17 publications we definitely succeeded our planned output. Also, SELMA organized two user meetings (User Board and User Day) which were well received with together more than 100 participants. The participation in the EU Innovation Radar showed that the project is also getting attention at EU level: All our 3 submissions were selected to be showcased on the Innovation Radar websites.

Enhancing the SELMA project and exploiting its results is made easier by gathering feedback at an early stage of the research and development process. Evaluating SELMA output both within the project (see D5.2 Interim Evaluation Report) as well as reaching out to potential users on conferences and workshops has been proven very important. In addition, SELMA also has set up the User and Advisory Board which serves as a sounding board for the developments. The impressions we received from the virtual User Group meeting in March and the first User Day Meeting in Bonn in September were already very promising.

Setting up the IPR framework for a successful exploitation of the SELMA results will be an important activity of the final year.

This is the second of three iterations. The final version is due at the end of the project (M36).