Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D1.2 Initial Prototype Report

Work Package	1
Responsible Partner	Priberam
Author(s)	Sebastiao Miranda, Guntis Barzdins
Contributors	
Reviewer	Andreas Giefer
Version	V1.0
Contractual Date	31 March 2022
Delivery Date	31 March 2022
Dissemination Level	Public

Version History

Version	Date	Description
0.1	21/03/2022	Initial Table of Contents (ToC)
0.2	29/03/2022	Internal Review version
0.3	30/03/2022	Finalization
1.0	31/03/2022	Publishable version

Executive Summary

SELMA's NLP research on transfer learning, user feedback learning and stream learning is being applied into three main use cases, Multilingual Media Monitoring (UC1), Multilingual News Content Production (UC2) and SELMA NLP Service Orchestration (UC0), through the development of different software prototypes.

This document provides an overview of the prototype development related to each of the three main SELMA use cases, including the status of the implementation of the requirements previously listed in D1.1.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

SELMA's NLP research on transfer learning, user-feedback learning and stream learning is being applied to three main use-cases, Multilingual Media Monitoring (UC1), Multilingual News Content Production (UC2) and SELMA NLP Service Orchestration (UC0), through the development of different software prototypes.

The UC1 requirements as described in D1.1 cover broadly the requirements of a Media Monitoring platform which Priberam is taking to market, named *Monitio*. In SELMA, we are leveraging the efforts from the commercial *Monitio* platform, allowing us to focus on the NLP research aspects of the Media Monitoring problem, mainly the activities related to Natural Language Processing, Transfer Learning, User-Feedback Learning and Stream Learning and also the activities related to performance scalability for processing massive streams.

A similar approach is being used on the development of the *plain X* platform for UC2, where the more commercial aspects are being financed jointly between Priberam and Deutsche Welle.

Use cases UC1 and UC2 are being developed by integrating SELMA's research output into two commercial software platforms: *Monitio* (UC1) and *plain X* (UC2). Use-case UC0 is a third use case meant to build and share an Open-Source platform core, incorporating SELMA research results which will allow the community to build stream processing software pipelines. This platform core is also currently in use in *plain X* (UC2), as well as being integrated in *Monitio* (UC1) – by applying the SELMA platform core to these two different use cases we're also proving that it's generic enough to be useful for the Open Source community in other NLP applications/products.

2. Multilingual Media Monitoring (UC1)

Within Use Case 1 (UC1), we are integrating results from the SELMA research tasks into the *Monitio* product, a Media Monitoring platform under development by Priberam, available at https://app.monitio.com. Some of these research results are improved models using transfer and stream learning (see D2.1 for technical details), while others are changes to the platform to allow learning from user feedback.

Another major change underway is integrating the SELMA orchestration platform as *Monitio*'s job orchestrator, which will allow *Monitio* to scale (see D4.1 and D4.2).

In this section, we present specific instances of SELMA research applied to *Monitio's* components, whereas in Section 5 we report a complete status of the requirements defined in D1.1.

2.1 Improvements on Multilingual News Clustering

The Multilingual News Clustering component has been improved through transfer learning by developing a new model capable of leveraging pre-trained crosslingual sentence embeddings in 50 languages *(en, ar, bg, ca, cs, da, de, el, es, et, fa, fi, fr, fr-ca, gl, gu, he, hi, hr, hu, hy, id, it, ja, ka, ko, ku, lt, lv, mk, mn, mr, ms, my, nb, nl, pl, pt, pt-br, ro, ru, sk, sl, sq, sr, sv, th, tr, uk, ur, vi, zh-cn, zh-tw).* See deliverable D2.2 for release details.

In the *Monitio* Platform, the results from this model are directly visible in the "Storylines" dashboard page, which shows the most relevant aggregated news stories within a specific time range and after applied filtering. See Figure 1.

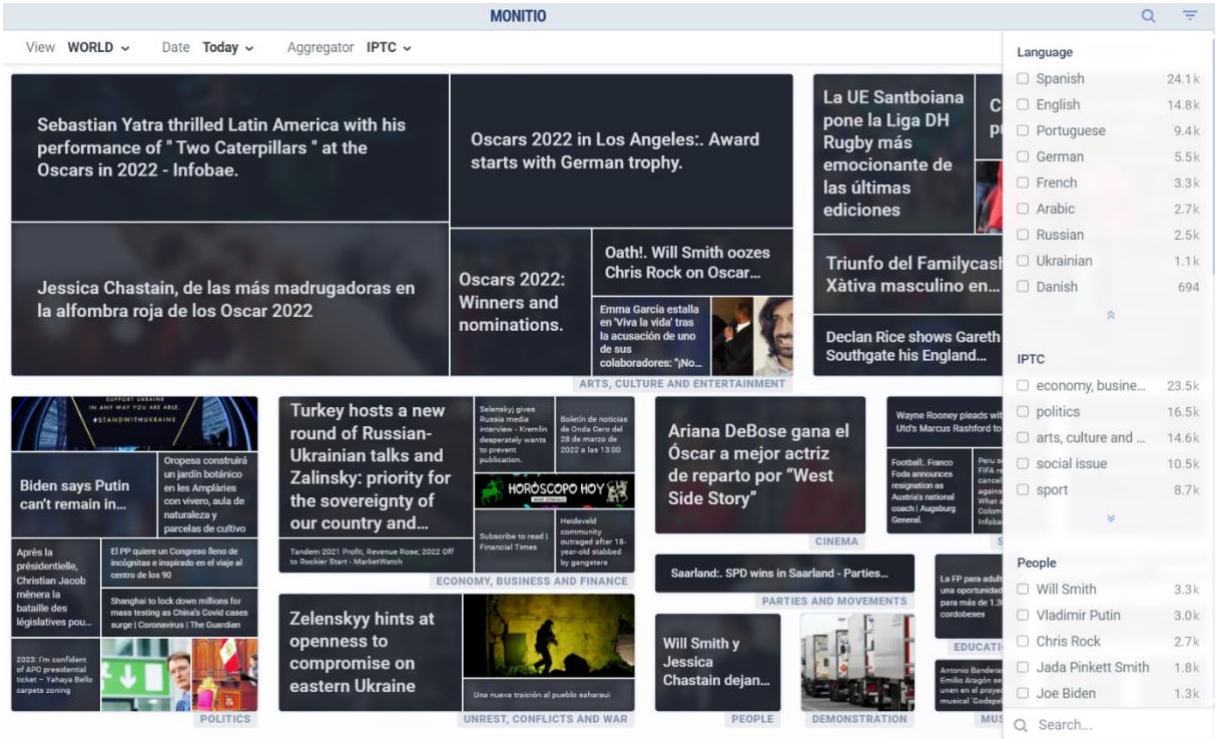

Figure 1 *"Storylines" dashboard from the Monitio platform, showing the clusters from the cross-lingual clustering model developed in SELMA*

2.2 Improvements on Multilingual Topic Detection

In the *Monitio* Platform, the results from this model are directly visible in the "Document" page, as seen in Figure 2 (IPTC Topics). These topics can also be used to filter documents on other pages, such as the "Storylines" page (see Figure 1).

Figure 2 *IPTC Topics detected on a Russian document. The tagging was done in the original language by a multilingual model, wereas translation is only used for showing the result to the user*

2.3 Document Tagging based on User Feedback

One application of User-Feedback-based models in *Monitio* is allowing users to create custom document tags. Users can associate documents with each custom tag (positive examples) and also give negative examples (teach *Monitio* that a certain document shouldn't be tagged with a certain tag). Over time, users' feedback is used to improve the model's output. This model is still early in development, but is already linked to a feedback interface. In Figure 3, we show

an example of adding a tag to a document. In this case, the user is creating a custom tag related to Oil Markets.

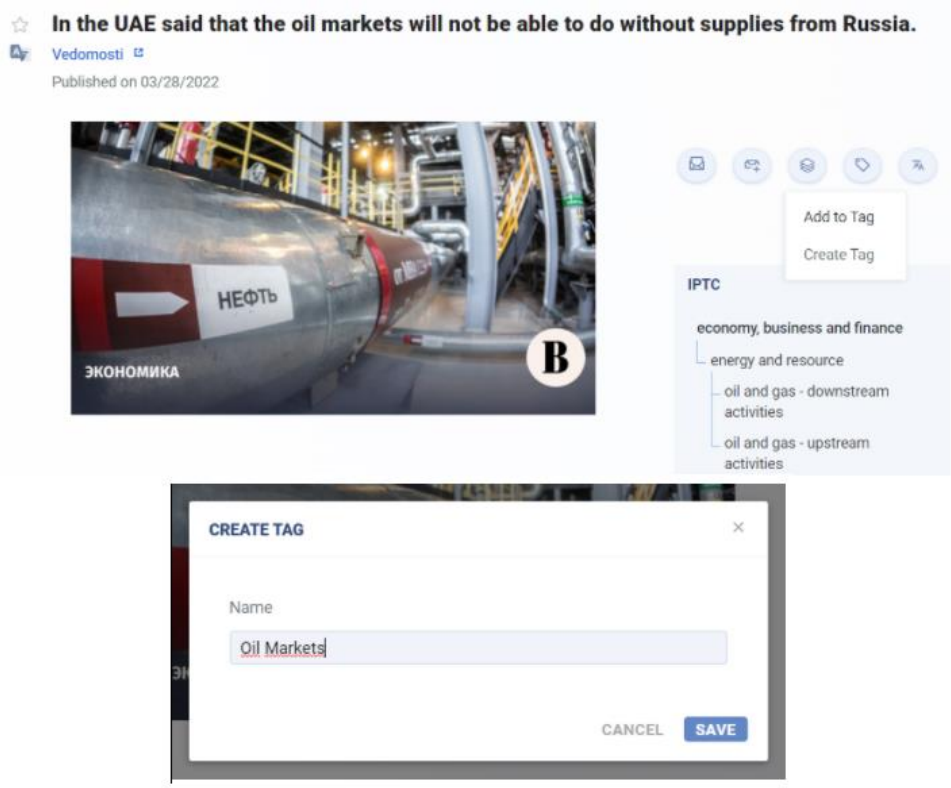

Figure 3 Adding custom, user-defined tags to documents

Figure 4 shows an integration diagram of this feature, called “Smart Tags” in *Monitio*. As mentioned, this system is still under development, but the first idea is, for each user, to create one tagger per different tag using a very light SVM model. After we have this baseline

implemented, other approaches to consider are few-shot transfer learning and also investigate active learning techniques.

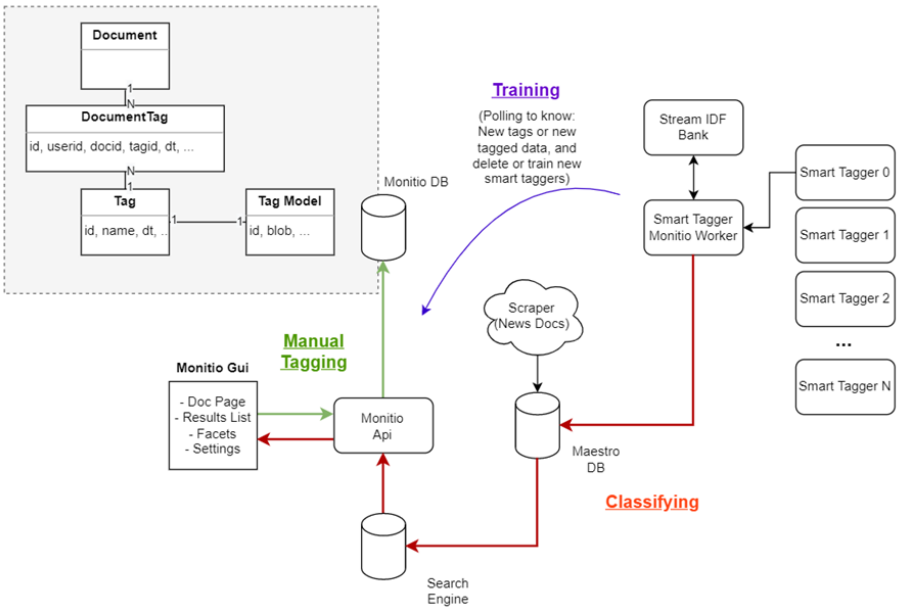

Figure 4 *Pipeline illustrating how custom, user-defined tags are associated with documents*

2.4 Entity Correction using User Feedback

We've prepared the *Monitio* platform to receive user feedback by allowing it to store user-edits of the output of some models. The first application we're approaching is allowing users to give direct feedback to the entities tagged and linked to wikipedia/wikidata. In this case, a user can go to the Document page in *Monitio*, select an entity and change the knowledge base link that the system has chosen (See Figure 5). These data can be leveraged in two ways:

(1) The first way, already implemented, is to use this feedback directly as string-match-based rules to fix future entity predictions. This works in some scenarios but it's not a scalable approach.

(2) The second way, under plan, is to integrate the user edits as a training signal in the entity linking model.

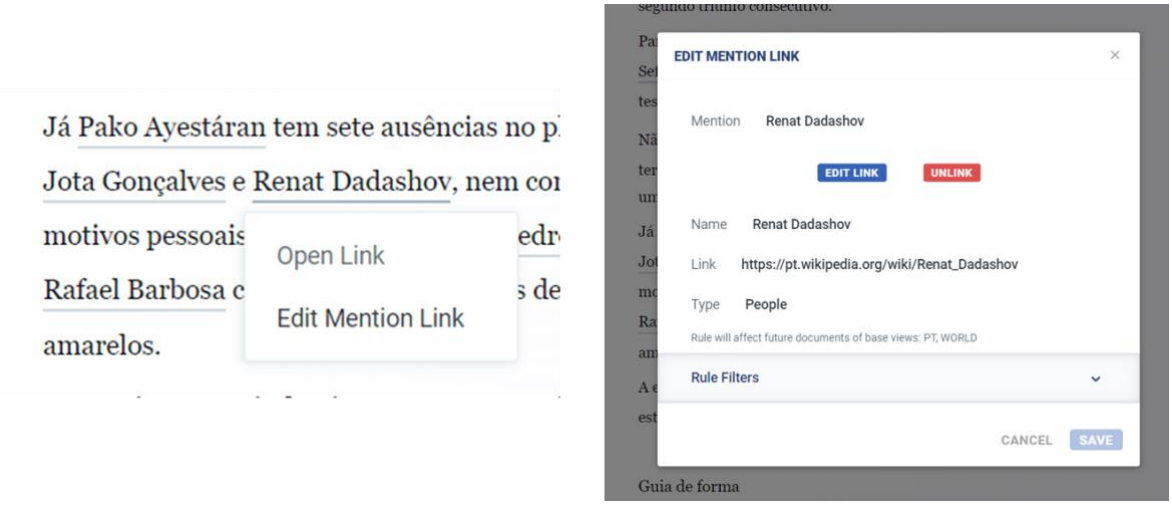

Figure 5 User feedback collection UI for entity linking

3. Multilingual News Content Production (UC2)

Within Use Case 2 (UC2), we're integrating results from SELMA into the *plain X* product, a Multilingual News Media Content Production platform under development by Priberam and Deutsche Welle, available at https://app.plain-x.com.

To achieve SELMA's research goals on learning from user feedback, *plain X* is being built having in mind storage and, later, serving of the original (automatic) NLP output and the corresponding user-edited versions.

plain X is using the SELMA orchestration to schedule and execute NLP jobs (see D4.1 and D4.2). In the case of *plain X*, the SELMA orchestration allows to execute NLP jobs not only from self-hosted APIs (e.g., the ones developed within SELMA) but also from many cloud providers (Azure, Google, etc).

3.1 *plain X* user interface

The *plain X* user interface is being developed to meet SELMA's UC2 requirements (see section 5). In this section we provide an overview of some of the most important views of the platform.

The entry point of the platform is the Library, a place where a user can find items they wish to work on (transcribe, translate, subtitle or voiceover). The Library can be filtered by Repository or source language. See Figure 6.

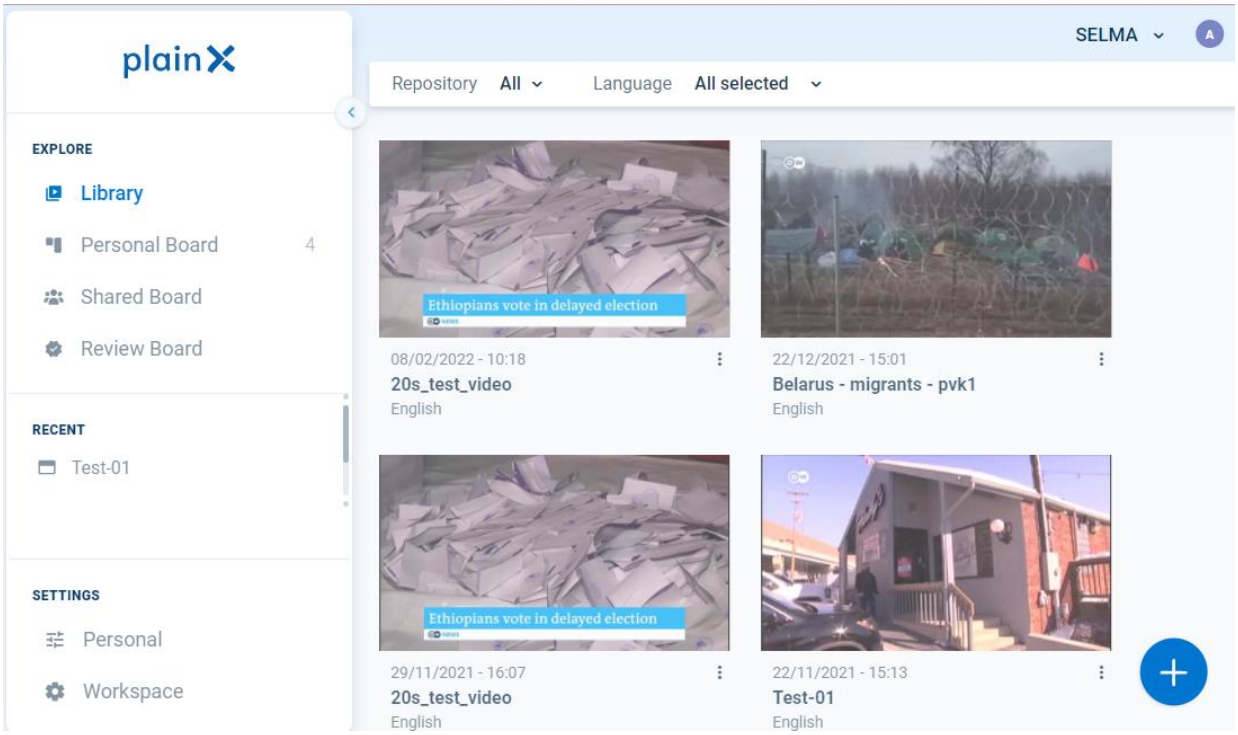

Figure 6 *plain X library showing available videos to work on*

Besides allowing to store and find video items to work on, *plain X* also allows one person or a team of people to coordinate their tasks using a task board, as seen in Figure 7.

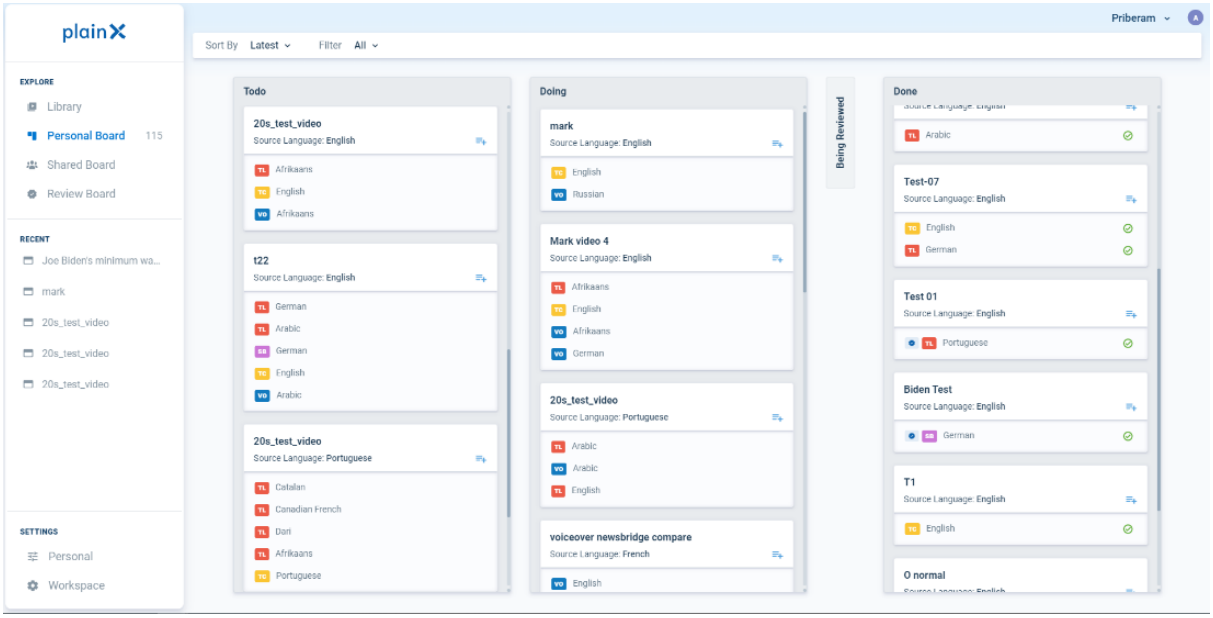

Figure 7 *plain X* board page

For each editing task, *plain X* offers a specific task page. In Figure 8, we show the Transcription task page which starts by showing to the user an automatic transcription and then allows them to edit this transcription. User edits are saved for future model improvement.

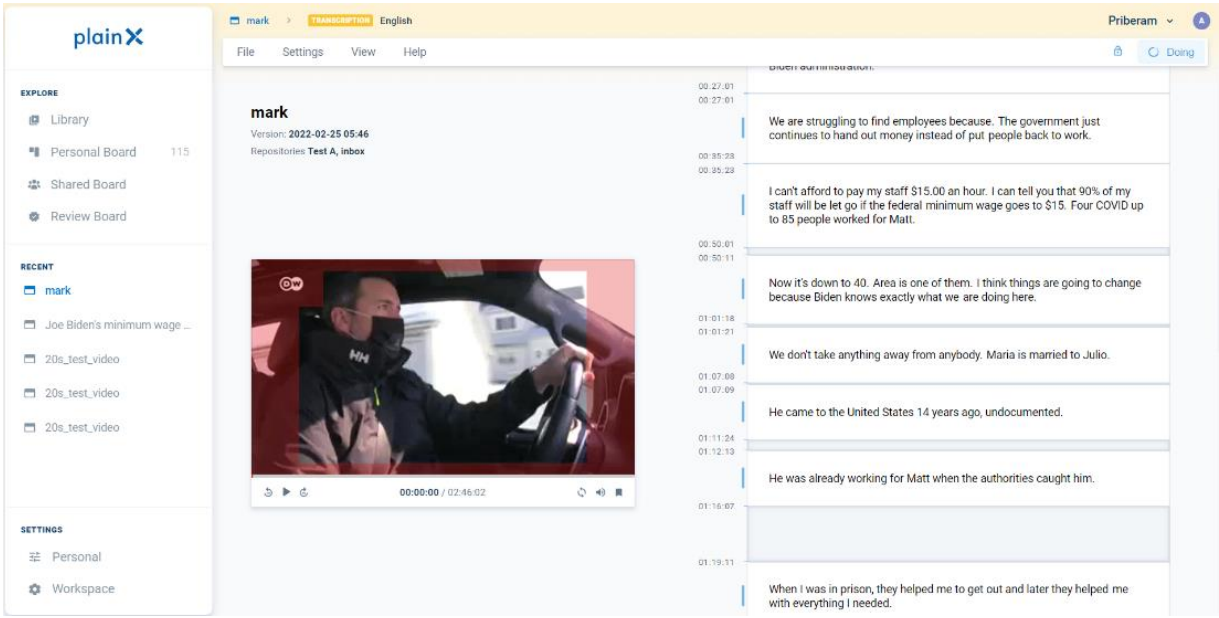

Figure 8 *plain X* transcription task page

In Figure 9, we show the Translation task page which starts by showing to the user an automatic translation and then allows them to edit this translation. User edits are saved for future model improvement.

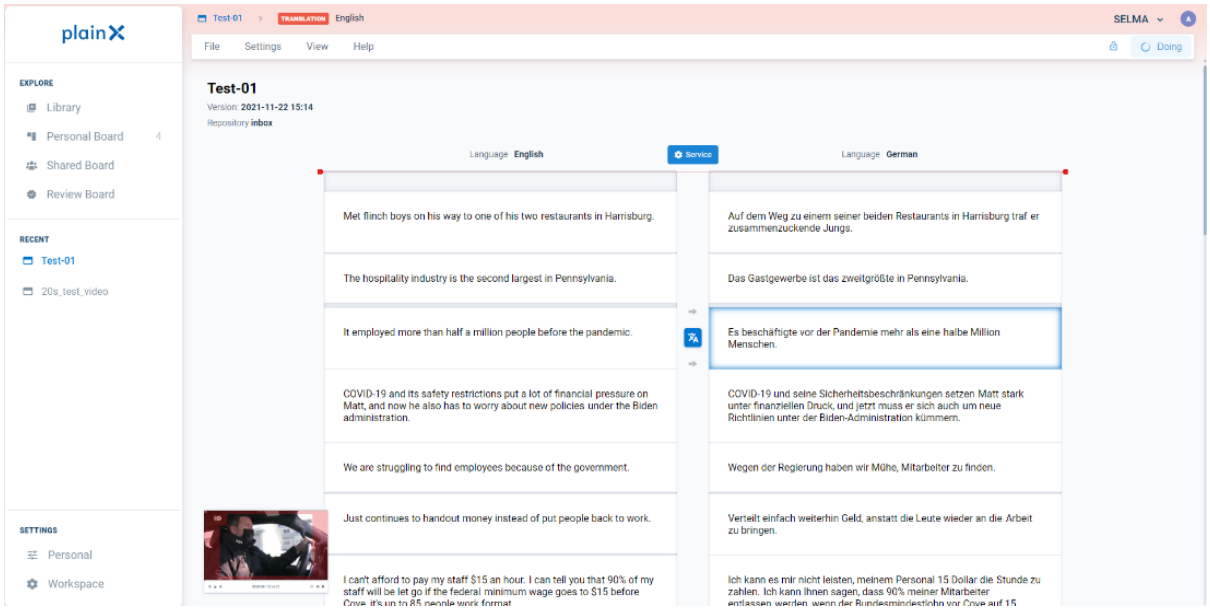

Figure 9 *plain X Translation task page, showing a video transcript in the original and target language*

Finally, in Figure 10, we show the Voice-over task page. A first automatic synthetic voice is generated and merged to the video's audio track. The user can customize the synthetic voice

generation by using a GUI tool to select text segments and change volume, pitch, pronunciation, etc.

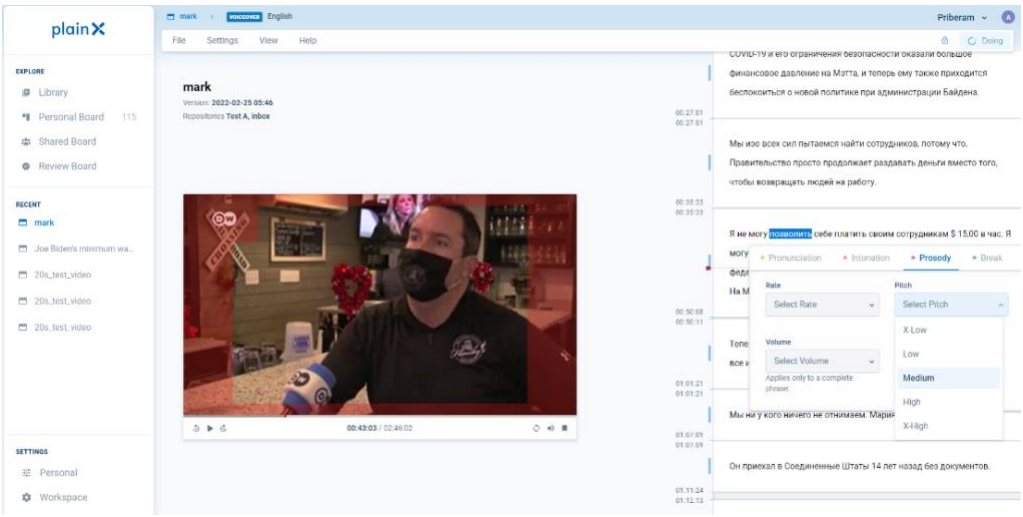

Figure 10 *plain X Voice-over task page, showing the GUI component to highlight text segments and customize synthetic voice generation parameters*

plain X also allows the creation and management of Users, Teams and Repositories (i.e. the Folders & Sources of news items). Another parameter which can be customized is which default engines should be used for each language / pair, as shown in Figure 11.

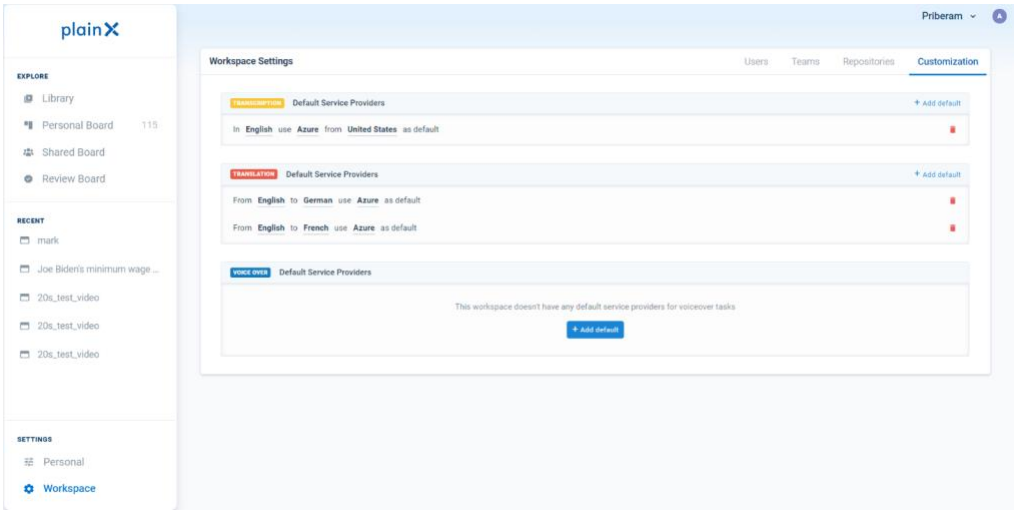

Figure 11 *plain X Settings page which allows customizing the workspace in terms of preferred NLP engines for certain language pairs*

4. SELMA NLP Service Orchestration (UC0)

Use-Case 0, also referred to as “SELMA NLP Service”, is a general open-source platform which serves as the core of UC1/2 and can be used by the community to implement other NLP platforms and products.

The SELMA Basic Testing and Configuration Interface (UC0) is an open-source software (https://github.com/SELMA-project/SELMA-project.github.io, (see Figure 12) for testing, deployment, scaling and monitoring of the NLP services developed within SELMA work packages WP2 and WP3. The NLP worker deployment utilizes a TokenQueue mechanism (described in D4.1) to deliver highly scalable **SELMA NLP Service Orchestration** for the primary Use Cases UC1 and UC2, as described in the following sections.

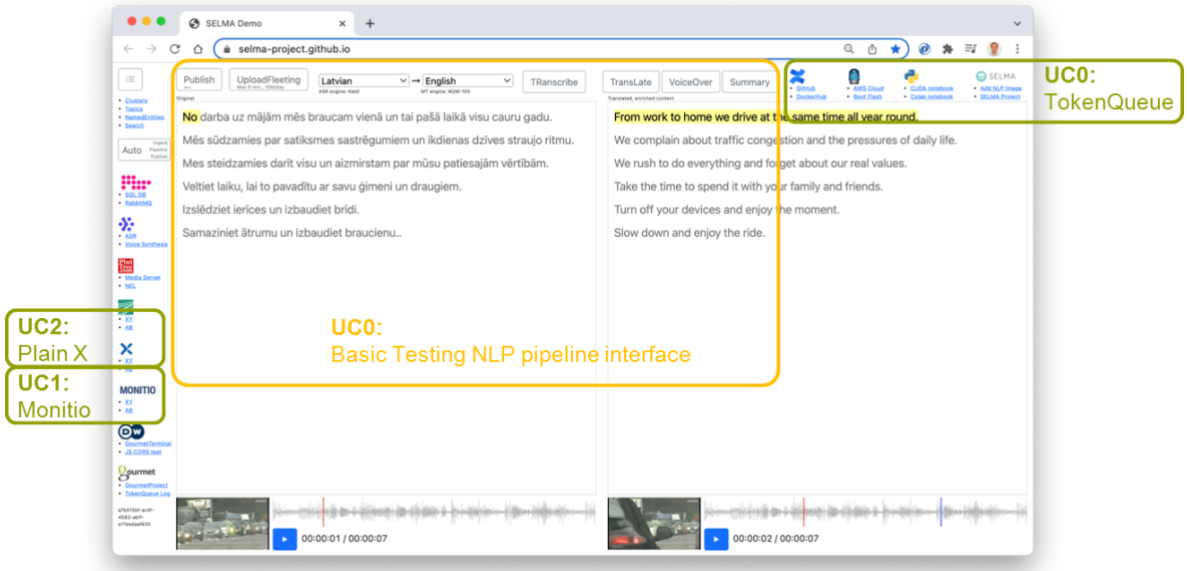

Figure 12 *SELMA Basic Testing and Configuration Interface (UC0)*

UC0 integrates with the NLP-pipeline execution orchestrator Maestro (described in D4.1) which is shared with UC1 and UC2. Consequently, the Maestro Orchestrator serves as a gateway between all three Use Cases UC0, UC1, UC2 allowing them to share the same NLP worker pool defined in the TokenQueue. In terms of control flow integration, Maestro Orchestrator uses TokenQueue via a REST API service.

5. External Use Cases: LETA

5.1 PiniTree: Rule-based Stream Learning for NEL

SELMA partner IMCS, University of Latvia, has been involved in the Named Entity Linking (NEL) topic for several years (Barzdins, 2020; Paikens, 2016a), jointly with the Latvian national news agency LETA and PiniTree.com startup. This has resulted in the development of the commercial PiniTree.com ontology editor. The latter integrates rule-based Stream learning of Named Entity Linking aliases as part of the entity database, against which the Named Entities are being linked. The PiniTree editor is one of the tools being integrated into the SELMA Platform and besides the LETA use case, it is available for wider exploitation along with other SELMA components. Within the initial release of the SELMA Platform, PiniTree is integrated in Use Case 0 as the backend content management system accessed via the "Publish" button.

5.2 Description of the NEL Stream Learning process in the PiniTree Editor

IMCS is collaborating with the Latvian national news agency LETA which maintains a database of the nationally significant organizations and persons. The PiniTree button in LETA's content management system allows one to link news articles to this database.

By pressing the PiniTree button, the current LETA news article opens in the PiniTree system which automatically colors the mentions of the significant organizations and persons found in the LETA database (Figure 13). The brown-colored names will mostly be recognized correctly due to the continuous stream-learning of aliases for the entities stored in the database. However, they also need to be validated by the user so that they become valid (green-colored) mentions of an organization or person.

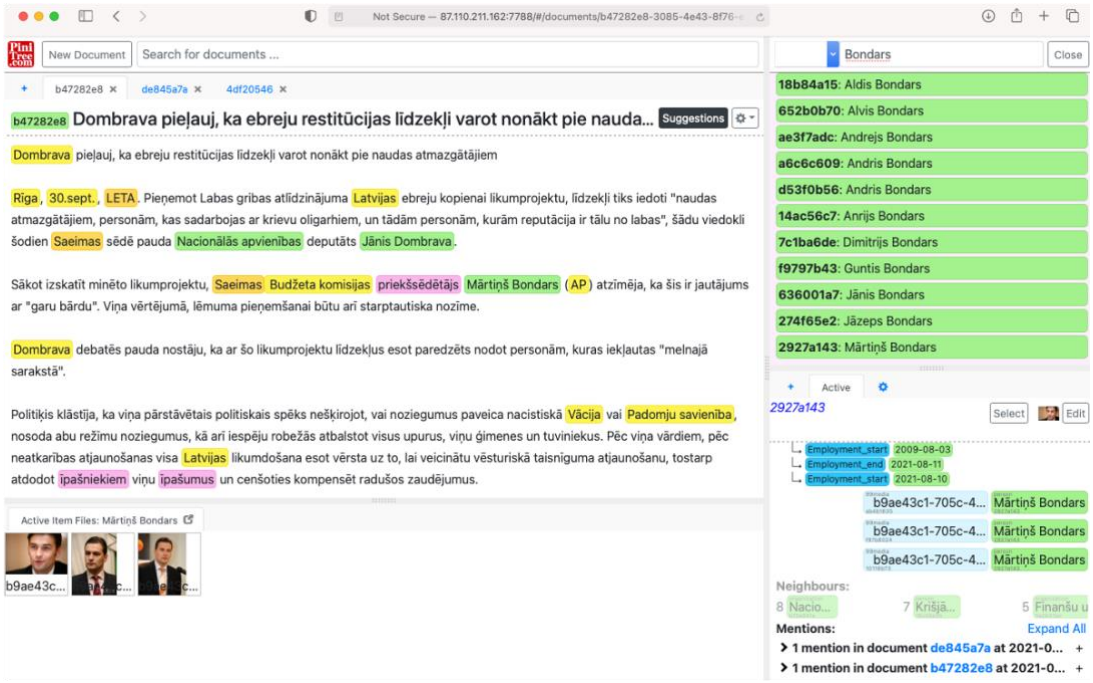

Figure 13 *PiniTree ontology editor includes NEL stream learning functionality*

If there are multiple persons with the same name or alias in the LETA database, they will be colored red, and the journalist must manually disambiguate the right person or organization and only then she or he will be able to validate the entity.

The article could also mention persons or organizations which are not yet included in the LETA database, they are colored in yellow by the Named Entity Recognition (NER) neural engine (Znotins, 2018) based on the massive LV-BERT large language model (Znotins, 2020). By clicking on the yellow-colored person, all the information about the person to be added is automatically filled in in the template on the right side of the window by the rule-based Latvian Part-Of-Speech (POS) and inflection engine (Paikens, 2016b). If everything is right, the journalist can take action so that the new person/organization is added to the LETA database. This is how the rule-based Stream learning is implemented in the PiniTree ontology editor powering the LETA database – the newly added entity (person, organization) description includes also all possible spelling aliases for the given entity which will be automatically matched to spot that entity in all further documents automatically.

To disambiguate between the persons and organizations with similar names, LETA widely uses the Firmas.lv database containing facts relating persons to the organizations where they are

owners or key employees. This way, the LETA database indirectly ingests up-to-date information from the Register of Enterprises of the Republic of Latvia while the PiniTree editor makes it easily accessible and allows to supplement the LETA database with new facts mentioned in the news articles.

When selecting an organization or person that is included in the Firmas.lv database in the news article, the most important facts about it are displayed automatically on the right side of the screen. By clicking on the displayed facts, one can navigate through the Firmas.lv data spider. The navigation history is displayed in the "History" field which allows the user to return easily to any of the steps visited earlier.

In news, new facts often appear that are not part of the Firmas.lv data. PiniTree allows journalists to add such facts to the LETA database. Violet-colored phrases indicate potential new facts mentioned in the document; they are recognized in the rule-based manner based on the LETA ontology (Figure 14). Clicking on the violet phrase opens a template for the new fact, where the roles mentioned in the fact must be filled out manually before adding the new fact to the LETA database by clicking the "Create" button.

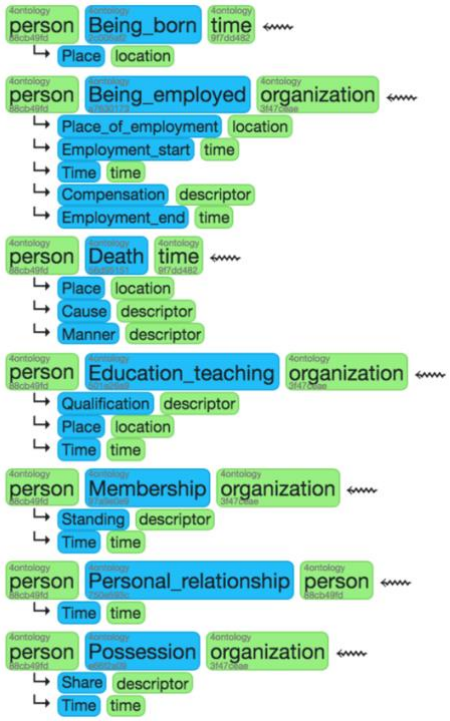

Figure 14 *LETA ontology used in the PiniTree ontology editor*

A new fact can also be created manually by marking the corresponding phrase in the document and pressing the "+" tab on the right side of the screen. In this case, all fields must be filled-in manually according to LETA's ontology scheme (Figure 15) which focuses around the concept of a person. The ontology scheme also allows adding fine-grained sub-facts to a given fact, such as "Location", "Time", "Qualification".

The "Edit" button can be used to edit information about organizations and persons stored in the database, such as adding a description, alternative names or a profile picture. Also, by pressing the document identification button at the upper left corner of the screen, facts or images can be added to the document as a whole, similarly to how they were added to a person/organization. In order for the newly added images to appear alongside the document, the PiniTree page must be reloaded in the web browser. It is possible to create also a visual mention for entities visible in the images by pressing the "Select" button and marking the corresponding region in the image. Regular and visual mentions appear as links under the entity at the right side of the screen.

6. Requirement Implementation Status

In this section we present the list of requirements as defined in D1.1 and the corresponding status in the UC1 and UC2 use cases as:

- Yes: Requirement Implemented
- Ongoing: Requirement Implementation Started
- Not yet: Requirement not Started

6.1 User & Platform Requirements

	Requirements	Use Case	MoSCoW	UC1 Status	UC2 Status
1	The System allows content to be ingested via standard interfaces as used by news organizations where available	1	Must	Ongoing	-
2	It is possible to use APIs to add content to the SELMA platform	1, 2	Should	Ongoing	Ongoing
3	The System ingests on-demand AV content in MP4 format	1,2	Must	Not yet	Yes
4	The System monitors selected social media feeds	1	Should	Ongoing	-
5	The System scrapes news articles from websites	1	Must	Yes	-
Module Functionality					

6	The system processes content in the 30+ SELMA languages	1, 2	Must	Ongoing	Yes
7	The system transcribes from audio	1,2	Must	Not yet	Yes
8	The system provides statistical analysis of ingested material	1	Must	Yes	-
9	The system provides automated translation of all SELMA languages into English as default	1	Should	Ongoing	-
10	The system provides automated translation into other SELMA languages upon request	1, 2	Should	Not yet	Yes
Output					
11	The system offers the possibility to create dashboard user interfaces	1	Should	Ongoing	-
12	The system provides a summarization of individual media items in original language and/or English	1	Must	Ongoing	-
13	User selects ingested channels to monitor	1	Must	Yes	-
14	User subscribes to notifications when relevant content arrives	1	Must	Ongoing	-
Content					

15	The system uses clustering technology to group individual media items into over-arching high-level clusters	1	Must	Yes	-
16	The system provides a visualization which contains a list of all high-level news stories that are relevant to the user, according to the preferences they have set	1	Must	Yes	-
17	On detection of a high-level news story, the system provides a default name for this story based on the clustering technology parameters	1	Could	Yes	-
18	The system offers the user the ability to follow a specific story and subscribe to updates regarding that story	1	Could	Not yet	-
19	The system allows the user to select a high-level news story and view the individual media items that are relevant to it	1	Must	Yes	-
20	For each high-level news story, the system displays a timeline. The system indicates where each individual media item fits on that timeline	1	Could	Not yet	-
21	When viewing an individual media item, the user has the ability to link back to the over-arching high-level	1	Must	Yes	-

	news story to which it is related (in order to view the other media items related to that cluster)				
22	The individual media items for a cluster continue to accrue for a selected period of time. However, the user can indicate to the system that the cluster is no longer relevant before this time has elapsed	1	Could	Not yet	-
23	The system retains a record of clusters that it has identified along with an indication of how many media items were identified by the system	1	Must	Yes	-
Entity Requirements					
24	The system uses entity identification technology to group individual media items by entities	1	Must	Yes	-
25	For each selected entity, the system displays a timeline	1	Must	Yes	-
26	The system carries out entity identification (person, organization, locations and events) using the original language and/or English translation	1	Must	Ongoing	-
27	The system provides a list of identified entities that are relevant to the user,	1	Must	Yes	-

	according to the preferences they have set				
28	The system allows the user to highlight an entity identified by the system and view the individual media items that include the entity	1	Must	Yes	-
29	When viewing an individual media item, the user can link back to the entity to which it is related (in order to be able to view the other media items related to that entity)	1	Could	Not yet	-
30	The system retains a record of entities identified along with an indication of how many media items were identified as relating to those particular entities	1	Must	Yes	-
News Story Requirements					
31	The system uses preferences set by the user to detect news stories of interest to the user	1	Must	Not yet	-
32	For each cluster, the system displays a timeline. The system indicates where each individual media item fits on that timeline	1	Should	Not yet	-
Breaking News Alert Requirements					
33	The system provides breaking news alerts that will correspond to	1	Must	Not yet	-

	individual media items in accordance with the preferences set by the user				
34	A breaking news alert consists of a textual description of the associated media item in the original language and/or in English, along with some specified meta-data (such as date and time)	1	Must	Not yet	-
35	The user selects their preferences for the type of clusters for which they wish to receive alerts (this may be related to particular event or entity types)	1	Must	Not yet	-
36	The user selects the manner and frequency at which they receive event alerts	1	Must	Not yet	-
37	Breaking news alerts are as close to live as is technically possible	1	Must	Not yet	-
General Functional Requirements					
38	The system monitors all input sources selected by the user	1	Must	Yes	-
39	The user can turn English translation on or off	1	Must	Ongoing	-
40	It is possible to associate a user with their team in the System	1, 2	Must	Yes	Yes

41	It is possible to indicate the role of a user in the System	1, 2	Could	Not yet	Yes
42	A user can share a cluster or an individual media item with their team	1	Must	Not yet	-
43	Once a user has indicated that a particular news story or cluster is no longer relevant, individual media items relating to that entity or cluster can be removed from the user's view	1	Must	Not yet	-
44	The user can flag a particular individual media item, entity or cluster and its related media items and indicate that they wish to save them for future reference	1	Must	Not yet	-
45	The user has an option in the system where they can view individual media items, entities or clusters which they chose to save (alongside all the individual media items related to these)	1	Must	Not yet	-
Media Item Requirements					
46	The system provides a clear visual indicator as to the nature of an individual media item (social media, blog, website, AV etc.)	1	Must	Yes	-

47	The user can view the detail of an individual media item (when applicable)	1	Must	Yes	-
48	For an individual AV media item, the user views the video and its original transcription, a translation in the prespecified languages (where applicable) and the meta-data associated with the item	1, 2	Must	Not yet	Yes
49	For individual AV media items, the system supports a player and editor with tools to ’scrub’ through the video, rewind and download	1, 2	Must	Not yet	Yes
50	For individual AV media items, the system supports a player and editor with tools to select elements to ’clip’	1, 2	Could	Not yet	Not yet
51	It is possible for the user to ‘clip’ an individual AV media item by means of text selection from the transcript	1, 2	Could	Not yet	Not yet
52	For other media occurrences (i.e. textual), the user views the text, its translation in the prespecified language (where applicable) and any meta-data associated with the item	1, 2	Must	Yes	Yes
53	The system provides a ’confidence level indicator’ which will indicate how strongly an individual media item	1	Must	Not yet	-

	relates to the existing or suggested clusters				
User Preference Requirements					
54	It is possible to set up a set of default sources that will be frequently used by a particular team	1, 2	Must	Yes	Yes
55	The system contains a predefined list of sources by region	1	Must	Not yet	-
56	The user can specify entities of particular interest to them	1	Must	Yes	-
57	The user can choose their region of interest in the System	1	Must	Not yet	-
58	The user can prioritize countries of interest within their region of interest	1	Must	Not yet	-
59	The system contains a predefined list of regions and countries	1	Must	Not yet	-
60	It is possible for the user to set time parameters in the system using an indicator such as a time slider to indicate the time period in which they are interested	1	Must	Yes	-
61	In general, the system supports input of preferences in a number of ways: From predefined lists, using data being encountered in the system (the system	1	Must	Yes	-

	will create new entities, events etc.), using free-format text (i.e. search boxes)				
62	The system is adaptable and configurable to the user's preferences	1	Should	Yes	-
Administration Requirements					
63	The user can log into the system	1, 2	Must	Yes	Yes
64	The user can log out of the system	1, 2	Must	Yes	Yes
65	The system supports a super user account	1, 2	Must	Not yet	Yes
66	The system supports an administrative user (for account management)	1, 2	Must	Not yet	Yes
67	The administrator can create teams in the system	1, 2	Should	Not yet	Yes
68	The administrator has typical administrator rights including add new user, remove users, update user profiles as well as the ability to manage data held by the System	1, 2	Should	Not yet	Yes
Search Requirements					
69	It is possible for the user to conduct a search based on an entity or entities	1	Must	Yes	-

70	It is possible for the user to search based on the type of the media item (e.g. social media, AV etc.)	1	Must	Yes	-
71	It is possible for users to search based on event types	1	Won't	No	-
72	It is possible to take a screen shot (or still frame) from an individual AV media item (rights to be considered here)	1, 2	Could	Not yet	Not yet
73	Entity search is able to handle a range of variable spellings for the same entity	1	Must	Yes	-
74	It is possible to train the system in relation to alternative spellings for searches. For example, it should be possible to link to alternative spellings and indicate that they relate to the same thing	1	Won't	No	-
Input Source Requirements					
75	The system informs the user if a source stops broadcasting	1	Could	Not yet	-
76	The system informs the user if the frequency at which a channel is broadcast, changes	1	Could	Not yet	-
Trend Analysis Requirements					

77	The user has the ability to set their preferences in the system with regard to the types of trend analysis they wish to see in the system	1	Must	Yes	-
78	The system offers the user various options around trend analysis including maps incorporating hotspots, graphs and timelines showing hotspots	1	Must	Ongoing	-
79	The system utilizes saved clusters and media items (that the system has saved along with the number of media items) to conduct trend analysis over a period longer than one week	1	Could	Yes	-
Generate Voice-Over					
80	The system generates a voice-over on request for individual AV media items	2	Must	-	Yes
81	The user chooses whether the voice-over is performed in the original or in the selected translation language	2	Must	-	Yes
82	The system provides a list of available synthetic voices for the user	2	Must	-	Yes
83	The user can choose and/or change which synthetic voice is used for the voice-over	2	Must	-	Yes

84	The user can change the synthetic voice per segment	2	Must	-	Yes
85	The user can amend the voice-over output including phonetics, pauses and pitch	2	Could	-	Yes
Edit Transcription/Translation					
86	The user can edit the transcribed text	2	Must	-	Yes
87	The user can edit the translated text	2	Must	-	Yes
88	The user saves the edited versions of the transcribed and/or translated text	2	Must	-	Yes
89	The user changes the engine and perform the transcription again for the whole text and/or by segment	2	Should	-	Yes
90	The user changes the engine and perform the translation again for the whole text and/or by segment	2	Should	-	Yes
System Learning and User Feedback					
91	The system is trained by the user. The system offers the user a selection of news stories and individual media items relevant to a cluster and the user accepts/rejects them as necessary, thus training the system	1	Must	Not Yet	-

92	The system is trained by the user. The system offers the user breaking news alerts and the user accepts/rejects as necessary – training the system to meet the user's preferences	1	Must	Ongoing	-
93	The system learns from the user's corrections and apply them throughout the text	2	Must		Not Yet
94	The system applies corrections on different levels (current and future items) based on the preferences set by the user	1, 2	Should	Not Yet	Not Yet
95	The system learns from the user's corrections for entities	1, 2	Must	Ongoing	Not Yet
Diversity Detection					
96	The system identifies the binary gender associated with the author's name (if present) of an individual media item	1	Should	Not Yet	-
97	The system provides the number of times each binary gender is mentioned in the media items	1	Must	Ongoing	-
98	The system identifies the gender of the protagonist in each individual media item (if applicable)	1	Should	Not Yet	-

99	The system provides the number of times each binary gender is mentioned in each topical cluster	1	Must	Not Yet	-
100	The system provides a visualization of the gender analysis	1	Must	Not Yet	-
101	The system provides all information pertaining to the diversity data (gender, age, sexual orientation, country of citizenship, medical condition) found on the Wikidata entry of relevant public figures where applicable	1	Must	Ongoing	-
102	The system identifies the gender associated with the named entities of type person, even if the gender information is not available in Wikidata (or the entity is not in Wikidata at all)	1	Should	Not Yet	-

Table 1 *User & Platform Requirements*

6.2 Technical Requirements

	Technical Requirements	**Use Case**	**MoSCoW**	**UC1 Status**	**UC2 Status**
Platform - Orchestration					
P1	Orchestrates processing jobs on the data stream, automatically	1	Must	Ongoing	-

P2	Orchestrates processing jobs on the data stream, on user request	1,2	Must	Not yet	Yes
P3	Allows listening for job results via push notifications (e.g, web sockets)	1,2	Should	Ongoing	Yes
P4	Allows listening for job state changes (errors, job progress)	1,2	Must	Ongoing	Yes
P5	Allows consulting the state of a job request / jobs on an item via an API (e.g., REST)	1,2	Must	Ongoing	Ongoing
P6	Accepts new “job workflow” requests, which may entail running several jobs organized in a graph of dependencies, on an item	1,2	Must	Ongoing	Yes
P7	Orchestrated jobs are eventually applied, meaning, a job cannot be lost - it is either completed successfully or logs an error	1,2	Must	Ongoing	Yes
P8	Orchestration-related configuration changes happen without downtime	1,2	Could	Ongoing	Ongoing
P9	Resilience to the unavailability / downtime of specific workers. Jobs wait until the worker recovers	1,2	Should	Ongoing	Yes

P10	The system allows parallel jobs to execute on the same item if they can be run that way according to the orchestrated job graph	1,2	Could	Ongoing	Yes
Platform - Worker management					
P11	Processing workers use Docker (or equivalent) containerization for deployment	1,2	Must	Yes	Yes
P12	The system is prepared for kubectl-compatible deployment in Kubernetes clusters	1,2	Should	Not yet	Not yet
P13	The system manages the lifetime of different worker containers	1,2	Could	Not yet	Not yet
P14	The system scales the number of worker containers according to the corresponding task queue flux	1,2	Could	Not yet	Not yet
P15	Worker-related configuration changes (new workers, worker scaling, etc.) happen without downtime	1,2	Could	Not yet	Not yet
Platform - Replication and Sharding					
P16	The system allows replication at the worker level	1,2	Must	Not yet	Yes
P17	The system allows replication at the orchestration controller level	1,2	Should	Not yet	Not yet

P18	The system allows replication at the database level	1,2	Should	Not yet	Not yet
P19	The system allows sharding at the orchestration controller level	1,2	Should	Not yet	Yes
P20	The system allows sharding at the database level	1,2	Should	Not yet	Not yet
Component - Online News Classification and Clustering					
C1	For each ingested news item, the system attributes a cluster	1	Must	Yes	-
C2	For each ingested news item, the system attributes an IPTC topic	1	Must	Yes	-
C3	The system clusters documents in an online fashion, e.g., without having to revisit all past decisions	1	Must	Yes	-
C4	The system clusters documents natively in all 30 SELMA languages	1	Should	Ongoing	-
C6	The system leverages user feedback on clustering decisions to improve future decisions	1	Should	Not yet	-
Component - Summarization					
S1	For each ingested news item, the system generates a summary	1	Must	Yes	-

S2	The system generates summaries natively in all 30+ SELMA languages	1	Should	Not yet	-
S3	The system generates summaries either from original text article or video transcripts	1	Should	Not yet	-
S4	The system leverages user feedback on summarization results to improve future summaries	1	Should	Not yet	-
Component - Machine Translation					
M1	The system translates a textual document by demand	1,2	Must	Not yet	Not yet
M2	The system translates a video by demand	1,2	Must	Not yet	Yes
				-	-
M3	The system translates between all 30+ SELMA languages	1,2	Should	Not yet	Yes
				-	-
Component - Automatic Transcription					
R1	The system automatically transcribes an ingested video or audio file	1,2	Must	Not yet	Yes
R2	The transcription is enriched by punctuation	1,2	Must	Not yet	Ongoing

R3	The transcription is enriched by speaker information	1,2	Should	Not yet	Not yet
R4	The transcription is enriched by named entity labeling	1,2	Must	Ongoing	Not yet
R5	The system transcribes all 30 SELMA languages	1,2	Should	Not yet	Yes
Component - Entity Recognition and Linking					
E1	For each ingested news item, the system detects named entities	1	Must	Yes	-
E2	For each ingested news item, the system links named entities to a knowledge base	1	Must	Yes	-
E3	The system links entities natively in all 30 SELMA languages, leveraging crosslingual representations	1	Should	Ongoing	-
E4	For each ingested news item, the system attributes a gender for each person named entity detected in the news item.	1	Should	Ongoing	-
Component - Story Segmentation					
G1	Each ingested long audio segment gets split into meaningful units	1,2	Could	Not yet	Not yet

G2	Speaker clustering is used to create speaker independent units	1,2	Could	Not yet	Not yet
G3	Speaker recognition automatically identifies the original speaker in each segment	1,2	Could	Not yet	Not yet
G4	Topic segmentation is used to separate by spoken content	1,2	Could	Not yet	Not yet
Component - Voice Conversion Synthesis					
V1	The text-to-speech system automatically produces voices in Latvian, German, and French	2	Must	-	Ongoing
V2	The text-to-speech system will be improved to better handle foreign words	2	Must	-	Not Yet
V3	A speech-to-speech translation system works at least on one language pair	2	Should	-	Ongoing
V4	A speech-to-speech translation system works at least on one language pair and can generate a synthetic voice in the target language close to the natural voice in the source language	2	Could	-	Ongoing

Table 2 *Technical Requirements*

7. Conclusion

This document presents the current state of the prototypes which address the three main use cases, Multilingual Media Monitoring (UC1), Multilingual News Content Production (UC2) and SELMA NLP Service Orchestration (UC0).

Within Use Case 1 (UC1), we're integrating results from the SELMA research tasks into the *Monitio* product, a Media Monitoring platform under development by Priberam, available at https://app.monitio.com. Within Use Case 2 (UC2), we're integrating results from SELMA into the *plain X* product, a Multilingual News Media Content Production platform under development by Priberam and Deutsche Welle, available at https://app.plain-x.com. The SELMA Open-Source Platform (UC0) is available at https://selma-project.github.io/.

The list of requirements from D1.1 is revisited and the corresponding progress reported.

Bibliography

Barzdins, G., Gosko, D., Cerans, K., Barzdins, O. F., Znotins, A., Barzdins, P. F., Gruzitis, N., Grasmanis, M., Barzdins, J., Lavrinovics, I., Mayer, S. K., Students, I., Celms, E., Sprogis, A., Nespore-Berzkalne, G., Paikens, P. (2020b). Pini Language and PiniTree Ontology Editor: Annotation and Verbalisation for Atomised Journalism. *In: ESWC 2020 Satellite Events. LNCS, Volume 12124, pp. 32-38.*

Peteris Paikens; Guntis Barzdins; Afonso Mendes; Daniel Ferreira; Samuel Broscheit; Mariana S. C. Almeida; Sebastiao Miranda; David Nogueira; Pedro Balage; Martins, Andre F. T. (2016a). SUMMA at TAC Knowledge Base Population Task 2016, *DOI: 10.5281/zenodo.827317*

Znotiņš, Artūrs & Barzdins, Guntis. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. *Baltic HLT, IOS Press, pp. 111-115, DOI 10.3233/FAIA200610.*

Znotins A, Cirule E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. *In: Human Language Technologies - The Baltic Perspective. vol. 307. IOS Press; 2018. p. 183–189.*

Paikens P. (2016b). Deep Neural Learning Approaches for Latvian Morphological Tagging. *In: Baltic HLT; 2016. p. 160–166.*

João Santos, Afonso Mendes and Sebastiao Miranda, “Simplifying News Clustering Through Projection from a Shared Multilingual Space” in Text2Story at ECIR, 2022.