Research and Innovation Action (RIA)   H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D6.5 Final Data Management Plan

Work Package	6
Responsible Partner	IMCS
Author(s)	Normunds Grūzītis (IMCS), Guntis Bārzdiņš (IMCS), Afonso Mendes (Priberam), Kay Macquarrie (DW)
Contributors	Andreas Giefer (DW), Yannick Estève (LIA), Peggy van der Kreeft (DW), Tugtekin Turan (FhG), João Prieto (Priberam)
Version	1.0
Contractual Date	31 March 2024
Delivery Date	28 March 2024
Dissemination Level	Public

Version History

Version	Date	Description
0.1	22.02.2024	First D6.5 draft
0.2	29.02.2024	Updated draft
0.3	07.03.2024	Integrated contributions from all partners
0.4	07.03.2024	Internal review
0.5	11.03.2024	Addressing review comments
0.6	21.03.2024	Pre-final version
1.0	25.03.2024	Final D6.5 version for submission

Executive Summary

The Data Management Plan provides an analysis of the main elements of the data management policy that is used by the SELMA consortium with regard to the datasets collected for or generated by the project. It addresses issues such as collection of data, data set identifiers and descriptions, standards and metadata used in the project, data sharing, property rights and privacy protection, risk mitigation, data long-term preservation and re-use, complying with national and EU legislation.

SELMA's central concept is to build a deep-learning NLP platform that trains unsupervised language models, using a continuous stream of textual and video data from media sources and make them available in a user/topic-oriented form in over 30 languages.

The knowledge learnt in the form of deep contextual models is transferred to a set of NLP tasks and made available to users through a **Media Monitoring Platform** (Use Case 1, UC1) to be able to handle up to ten million news items per day. The media monitoring platform is able to transcribe, translate (on demand), aggregate, write abstractive summaries, classify, and extract knowledge in the form of entities and relations and topics and present all this to the user using new visualizations and analytics over the data. The learnt contextual models are also applied to a **News Production Platform** (Use Case 2, UC2), using enriched models for transcription (ASR) and translation (MT), providing journalists in an operational editorial environment with a multilingual tool that will be able to learn over time. For testing the NLP components and pipelines of the SELMA platforms and tools, a **SELMA Basic Testing and Configuration Interface** (Use Case 0, UC0) has been additionally introduced, which has evolved into the

SELMA Open-Source Software Platform. It is used as both an internal testing platform[1] and a public demonstration platform[2] of the SELMA components and pipelines.

[1] https://selma-project.github.io

[2] https://selma.ailab.lv

Table of Contents

Table of Figures

Table of Tables

1. Introduction

The Data Management Plan (DMP) functions as a central tool for risk mitigation associated with data protection within the SELMA project and regarding its outcomes: the media monitoring and news production platforms (UC1 and UC2), the SELMA open-source platform (UC0), and datasets released for the research community. The DMP includes the following aspects:

- It describes what research and innovation activities of the project use which data, and who is responsible for handling, storing, and destroying the data (data processing).
- It describes the purpose of SELMA's data-driven research and innovation activities and clarifies the substantial public interest in the project's results.
- It describes the data protection safeguards that are put in place.
- It identifies the countries in which data is processed or reside, together with an understanding of the national privacy and data protection regulations, and engagement with the relevant data protection agencies.

The DMP considers the personal data protection and copyright protection issues addressed in D8.1 Ethics Deliverable, including information flows in the project, identification of the privacy and related risks (particularly regarding monitoring data; see Section 2.6), and actions taken by SELMA to reduce the identified risks. The issues addressed here are also part of the ethics, project management, and evaluation reports.

This is the final version of the SELMA DMP – an update of its interim version (D6.3), specifying the final language coverage, data amounts and availability of the technology-specific datasets produced, used and released by SELMA.

2. Types of Data Collected

SELMA develops language data processing platforms and tools for dealing with large volumes of data across many languages and different media types. It has a range of technologies that are implemented using real-world language resources (text and audio/video) as training and evaluation data. These technologies include automatic speech recognition and synthesis, machine translation, speech translation, named entity recognition and linking, text classification, clustering, spoken language understanding, text and speech summarization.

Data is collected in 30+ languages in which Deutsche Welle (DW) publishes content: Albanian, Amharic, Arabic, Bengali, Bosnian, Bulgarian, Chinese (Simplified and Traditional), Croatian, Dari, English, French, German, Greek, Hausa, Hindi, Hungarian, Indonesian, Kiswahili, Macedonian, Pashto, Persian, Polish, Portuguese for Africa, Portuguese for Brazil, Romanian, Russian, Serbian, Spanish, Tamil, Turkish, Ukranian, Urdu.

The project consortium includes two primary data providers. DW is an international broadcaster with a wide range of languages covered and is acting in the project primarily as a coordinator, user partner and content provider. Priberam is a Portuguese language technology company, and it has a double role in the project as a technology developer and a content provider. Data has also been provided by IMCS for Latvian for NER annotation.

The two primary use cases that put the data to use are:

- MONITIO – a **Media Monitoring Platform** (Use Case 1) for handling up to ten million story segments per day;
- *plain X* – a **News Production Platform** (Use Case 2) – a multilingual editorial environment for journalists.

Both DW and Priberam target the above use cases, where DW is testing and incorporating them in their production workflows, and Priberam is actively making them available for testing by selected clients.

Technically, Use Case 0 (UC0) – SELMA Basic Testing and Configuration Interface, which is maintained by IMCS, has also been introduced. It was initially used by the SELMA partners for testing the language processing components and pipelines of the SELMA platforms and tools. However, UC0 is open source[3] and publicly available[4] and therefore is increasingly used as a public SELMA demonstration platform. It should be noted that UC0 does not ingest data from external sources, but only uses the data provided by its users (testers), and this data is not stored after being processed, and no logging is done regarding user actions (in fact, even no user accounts are maintained for UC0).

"Collection of data" in this report refers to the acquisition of data by the consortium, primarily through content provision by DW and Priberam but also through language data provision by other SELMA partners for the development of the SELMA language processing components.

2.1 Data Types

Data collection can be grouped according to the following criteria:

- Intended use:
    - ingestion or monitoring data,
    - training data,
    - test data,
    - user data.
- Language processing technology:
    - speech recognition,
    - speech synthesis,
    - machine translation,
    - speech translation,
    - named entity recognition,
    - named entity linking,
    - text classification,
    - clustering,
    - text summarization,
    - speech summarization.
- Data type:
    - metadata,
    - text,

[3] https://github.com/SELMA-project/UC0-OpenSource

[4] https://selma.ailab.lv

    - audio & video.
- Delivery type:
    - batch data (incl. text streams, no audio live streams);
- Language:
    - 30+ DW languages
- Content and language data provider/user:
    - DW,
    - Priberam,
    - other partners.
- User personal data:
    - email,
    - password,
    - name.
- User feedback (e.g.)
    - user edited transcripts,
    - user platform usage,
    - user corrected named entities.

We divide data requirements and data provision into four major groups:

- Regular content that is ingested and analyzed for media monitoring.
- Specific training and testing datasets that are collected for the development of language technology components, i.e., various neural language models.
- User data which is needed to ensure restricted access to the MONITIO and *plain X* platforms is securely processed and stored to comply with GDPR.
- Public datasets referred to in literature, and datasets licensed by the research partners.

2.2 Requirements for Monitoring Data (UC1)

Use Case 1, the media monitoring platform, is a tool for monitors to investigate, relate, check content produced by media publishers. To perform its objectives, three main types of data are collected:

- Ingested media content from the web, enriched by applying various technologies.
- User data for authentication and verification of access levels.
- User data collected from user interactions with the software.

2.3 Requirements for News Production (UC2)

Use Case 2, the news production platform, is a multilingual tool for content adaptation: to create transcripts, translations, subtitles and synthetic voice-over of videos, audios or texts. It is designed to simplify and speed up the workflow, significantly reducing the time needed for the tasks. To perform its objectives, three main types of data are collected:

- Ingested media content by user request and the associated task data (transcripts, translations, subtitles and voice over).
- User data for authentication and verification of access levels.
- User data collected from user interactions with the software.

2.4 Requirements for the Open-Source Platform (UC0)

Use Case 0, the open-source platform, is a tool for testing and disseminating the language models developed or fine-tuned in the project. This tool does not collect any personal data and does not store data submitted for processing.

2.5 Requirements for Technology-Specific Data

Requirements and specifications for technology-specific datasets are gathered in WP2 and WP3, detailing what type of data and what quantities are needed. The SELMA partners have directly supported the technology development by providing the necessary training and test datasets for the various language processing components whenever possible. The provision depends on the availability of such data and on the required workforce for preparation and adaptation of the datasets. All SELMA partners realize that training and test data is needed to develop high-quality language processing components for the large variety of SELMA languages.

The members of the consortium are aware that the kind of artificial intelligence models produced are conditioned by the bias contained in the training datasets. It is the responsibility of each research partner to address and assess the quality of the datasets used. In the Ethics deliverable (D8.1), the SELMA consortium has produced an assessment of the output models, based on the datasets used for training, the risks, and their possible dual use.

All the 30+ languages are supported by one or another of the SELMA platforms, either by in-house development of the respective language processing components or by exploiting third-party APIs or language models. The focus in the scope of the project, however, is on a selection of high- and low-resourced languages: English, German, French, Portuguese (both versions), Spanish, Italian, Turkish, Dutch, Urdu, Bengali, Amharic, Tunisian, as well as Russian, Ukrainian and Latvian.

2.5.1 Raw Data and Metadata

For training state-of-the-art *wav2vec*-based speech recognition models for selected SELMA languages (English, German, French, Spanish, Russian, Portuguese), a large amount (at least several hundred hours) of diverse audio/video recordings are required for each language. Datasets of the specified amount were provided to the SELMA technology partners for the selected languages by DW.

To reuse the same datasets for training abstractive speech summarization models, the audio/video recordings have to be complemented with the corresponding text teasers. Such metadata already exists for a large part of DW audio/video recordings and were ingested via the DW API. Regarding additional quality criteria of audio/video recordings, no background music is required, shorter clips are preferred over longer ones.

For training and fine-tuning abstractive text summarization models, large amounts of news and their human-produced summaries are required. One of the standard benchmark datasets for abstractive summarization in English is the CNN/DailyMail dataset[5] which is available under an Apache-2.0 open source/data license. It contains ~300k news articles paired with human-written highlights. To acquire datasets of similar size (order of magnitude) for abstractive summarization in other selected SELMA languages, the DW API has been used.

For training entity representations, the Wikipedia and Wikidata public data dumps are used. Additionally, to take into account entity drift and being able to populate the knowledge base,

[5] https://huggingface.co/datasets/cnn_dailymail

an online mechanism is applied to update these representations. For this, we use all data we have from the scraping of news web sites. This data is stored on Priberam's internal servers. Researchers in the project can request or access this data using APIs with proper authentication schemas implemented to safeguard data access (an API key is provided to access the data; the API is implemented using the HTTPS protocol).

2.5.2 Transcribed Data

For the automatic speech recognition and transcription (ASR), in addition to the raw audio/video datasets, already existing datasets of transcribed speech corpora are also re-used in the SELMA project to fine-tune, for instance, *Whisper*-based ASR models for selected languages. No additional datasets of transcribed speech have been created for the ASR development in SELMA.

For the automatic text-to-speech synthesis (TTS), however, transcribed speech datasets of a limited amount have been created for selected DW languages (namely, Brazilian Portuguese and Urdu) for which appropriate existing datasets are not sufficiently available.

Thus, the ASR and TTS components are developed and integrated for the SELMA platform based on:

a) previously created datasets of transcribed speech, some of which are proprietary or otherwise restricted-access datasets but are available to the SELMA partners for internal use:

i. open-access datasets like the multilingual M-AILABS[6], CSS10[7] and CommonVoice[8] speech datasets, the TED-LIUM3 speech dataset[9], and the recent Spotify Podcast Dataset[10] have been considered

ii. restricted-access datasets like QUAERO and ETAPE, as well as a dataset for training Latvian TTS, are available for internal use only

b) previous and current work on acoustic and language modelling, and ASR / TTS system development (incl. third-party APIs) for the high-resourced SELMA languages

c) current work on transfer learning of acoustic and language models for targeting selected low-resourced priority SELMA languages

d) creation of relatively large audio/video datasets (nearly 16k hours of diverse recordings; see Table 1 in Section 2.7) for selected SELMA languages to develop pre-trained *wav2vec* models (in addition to prior audio/video datasets available to the SELMA partners for internal use)

e) creation of limited amounts of transcribed speech datasets for selected SELMA languages: at least 10 hours of transcribed single-speaker audio data is required per language to have a valuable training dataset for an end-to-end TTS system. DW provided for this purpose editorially corrected manuscripts/transcripts (for Brazilian Portuguese and Urdu).

For each audio file in a transcribed speech corpus, a correct verbatim transcription of the spoken text is required. For the speech summarization needs, however, a text teaser is required and provided instead of a full transcription.

[6] https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/

[7] https://github.com/Kyubyong/CSS10

[8] https://commonvoice.mozilla.org

[9] https://www.openslr.org/51/

[10] https://podcastsdataset.byspotify.com

Segmented and aligned data with timecodes is preferred, but data without timecodes is also useful, as timecodes can be added automatically.

The requested encoding for the transcripts (ASR / TTS) and text teasers (speech summarization) is UTF-8. The specific data format for each language is clarified between the data provider and the technical partners.

2.5.3 NER Annotated Data

For cross-lingual training of named entity recognition (NER) models, a multilingual NER-annotated dataset is required. For a selected subset of SELMA languages for which compliant prior datasets are not available, a representative set of documents (news articles), 50–750 per language were annotated by SELMA partners according to a common NER annotation schema. This was done semi-automatically with manual validation. Since the manual annotation process is very time-consuming, the whole annotation process was sped up by manually annotating 50–100 documents, then training a neural model for automatic NER, followed by a manual curation of automatically pre-annotated remaining set of documents (if any). The total required number of annotated/curated documents per language was gradually reduced for each new language due to the efficient language transfer mechanisms developed in the project (i.e., by joint training with more data of high-resourced languages and less data of low-resourced languages).

Regarding the document selection for each language, the focus was on news items and bulletins, i.e., broadcast news that is publicly available text data and is the scope of the project. This facilitates not only data collection but also sharing, since named entity annotation involves random personal data; in this case, data about public persons (mentions of person names and related entities). Nevertheless, the set of selected articles for each language has to be diverse (representative) in terms of topics, time periods, authors, channels. Therefore, datasets for NER annotation were partially collected from DW news feeds but were mixed with articles from other sources as well.

As for the common NER annotation schema, the Priberam Named Entities Annotation Guidelines (see Annex of D6.1) was used as the fundament and orientation for the SELMA multilingual dataset.

The multilingual dataset for training and evaluation of hierarchical NER systems would have a significant impact on the NLP research community, if it is released in the public or academic domain by the SELMA consortium. We are still seeking authorization from the original copyright content holders to publish the NER-annotated dataset with an open or academic license. Nevertheless, this would exclude the prior datasets annotated for Portuguese, French, English, Spanish and German that are used in the scope of the project but will not be released with an open license. For the rest of the languages, a license agreement is pending, and copyright infringement could be avoided by scrambling the datasets before releasing them (e.g. by randomly reordering the sentences in all the articles), however, this would narrow down the potential use cases of this dataset.

2.5.4 Entity Linking Data

For entity linking, SELMA uses open-access data from Wikipedia and Wikidata to train the base entity representations. To train and evaluate the disambiguation models, the project is using open-access datasets, namely: AIDA-CoNLL, Voxel, WikiMentions and TAC.

Additional automatic and human evaluations are done on UC1 monitoring data.

2.5.5 Topic Detection Data

For text classification, Priberam had previously acquired a dataset from the Portuguese News agency LUSA and recently licensed a dataset[11] from the Finish News Agency Archive. Both datasets contain news articles manually annotated with IPTC subject codes where the LUSA dataset from 2009 to 2015 and the Finnish dataset contains articles from 1992 to 2018. This new dataset together with the previously licensed LUSA dataset facilitates further exploration of the multilingual classification task and expands the diversity of the dataset.

[11] STT. Finnish News Agency Archive 1992-2018, source [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2019041501

Additional automatic and human evaluations were done on UC1 monitoring data.

2.5.6 Storyline Clustering Data

For the online clustering training and evaluation data, we reuse in SELMA the Cross-Lingual Document Similarity and Event Tracking dataset provided by Rupnik.[12]

Additional automatic and human evaluations were done on UC1 monitoring data.

2.5.7 Audio Datasets for Speaker Recognition and Diarization Systems

Speaker recognition models utilize vocal traits to identify individuals. Initially, we used the VoxCeleb dataset,[13] which is rich in diverse celebrity voice recordings over YouTube, to develop deep learning models for extracting unique features, known as speaker embeddings.

Next, we enhanced real-world performance using Mozilla's CommonVoice,[14] a multilingual collection. Integrating new speakers to our model, we measured the robustness and adaptability across varied languages and accents. This two-phase strategy ensured our speaker recognition system's high accuracy and generalizability.

For speaker diarization, which is segmenting audio by speaker identity, we chose VoxConverse.[15] Its wide array of audio scenarios challenges our models with different speaking styles, overlaps, and noise levels. Evaluation on VoxConverse confirms our system's precision in speaker separation, proving its effectiveness for transcription, multi-speaker analysis, and detailed audio processing.

[12] Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., Grobelnik, M. News Across Languages - Cross-Lingual Document Similarity and Event Tracking. In: Journal of Artifcial Intelligence Research, 55, Special Track on Cross-Language Algorithms and Applications, 2016.

[13] https://www.robots.ox.ac.uk/~vgg/data/voxceleb

[14] https://commonvoice.mozilla.org/en/datasets

[15] https://www.robots.ox.ac.uk/~vgg/data/voxconverse

2.6 Provision of Monitoring Data

Online data is continuously being collected and ingested into the SELMA platform for the media monitoring use case (Use Case 1). Audio and video data is being collected from Twitter/X and YouTube channels for selected media providers through the ingestion pipeline. The platform is ingesting about 300.000 news articles per day. Data from DW, covering all the 30+ SELMA languages is ingested into a specific MONITIO scenario.

Content is ingested into the SELMA platform using, depending on the source, one of the following methods (in the order of preference):

- RSS feeds
- API calls
- crawling XML site maps
- scraping document links from specific internet sites when none of the above possibilities are available.

In general, the most robust and flexible way to collect media content is via a combination of RSS feeds or XML sitemap ingestion and consequent site scraping (to get the full content of news items). In the case of DW, full content and metadata ingestion through the DW proprietary API is also done for a better data quality (in comparison to scraping); see Section 4.2 for more details.

DW content ensures testing the multilingual aspects of the SELMA platform, but the amount of data is not large enough for scalability testing. Data source diversity and large coverage is required for the actual monitoring use case, therefore monitoring data from many other public sources has been collected and ingested into the SELMA platform (UC1). News items from other public media sites are collected and provided by Priberam: by scraping news portal content based on XML site maps, by ingesting RSS, news sitemaps, sitemaps and by scraping links from specific sites. Since media companies are increasingly publishing unique content on social media platforms like X, Facebook, Instagram, TikTok and YouTube, we have successfully applied for access to gather data from public media pages on Facebook, Instagram and Twitter/X. We do not collect any personal data from social media users or any aggregated data (statistics) that could be used to quantify the reach of particular media items or media producers. We have arrived at the conclusion that the research being done in the scope of the project would not even benefit from aggregated user data from social media. As such, the adopted policy for collecting data from X (Twitter), Facebook and other social media is as follows: data is collected from public pages of media publishers keeping only the original published text; tweets, comments, replies and other user-generated content are not collected.

Additionally, we collect entity metadata from the open Wikidata[16] knowledge graph hosted by Wikimedia Foundation. For each processed article (a news item), the automatic named entity linker (NEL) of the SELMA platform (UC1) assigns a set of disambiguated entities (Wikidata identifiers of persons, organizations, etc.) and their Wikidata properties (e.g., binary gender and age; see Section 4.4 for more detail) based on the entity mentions in the article, which are detected by the automatic named entity recognizer (NER) of the SELMA platform (UC1). The linked entities and their properties are stored and indexed in a database of the UC1 platform. Thus, articles can be queried and retrieved by specific entities or common properties of named entities. This allows for aggregation and monitoring of some general diversity aspects (like binary gender, age) in media, but it does not allow for processing and analyzing any sensitive diversity aspects (like ethnicity, religion, education) since the SELMA consortium has abandoned its initial plan to collect such potentially sensitive personal data categories for the diversity monitoring. Nevertheless, it might still raise some potential ethical issues. To mitigate such risks, the SELMA project closely and regularly evaluated the development of the UC1 platform and has established a risk analysis and mitigation procedure with a focus on ethical and privacy concerns towards a potential use of the technology within UC1 and in the commercial platform. The ethical concerns and mitigation of risks are addressed in more detail in D8.1.

By design, UC1 (MONITIO) implements a comprehensive set of rules enabling it to comply with the stipulations of licensing agreements, as well as to deal with more generic copyright restrictions, such as the freedom to index content but only allowing partial display. To ensure

[16] https://www.wikidata.org/wiki/Wikidata:Main_Page

the respect of copyright, media monitoring content data is only shown to users of the platform when Priberam has established an agreement with the publishers or their representative associations, and in such a way as to respect the restrictions laid down in those agreements. Otherwise, to comply with copyright laws, for example, content may be indexed but only links and/or titles or excerpts of the articles displayed.

Currently, MONITIO ingests:

- Licensed content for Portugal (Visapress), Spain (Cedro) and UK (NLA);
- DW content;
- Other unlicensed content published by media publishers; this content is restricted on the platform: the users cannot see the text in accordance with copyright laws, and only a link to the original article is provided;
- Open access content;
- Entity specific, i.e. on demand/contractual content that may be composed by free content and/or entity owned, and/or entity licensed sources which are accessible only to that entity (client or partner).

Additionally, in the scope of SELMA research purposes, MONITIO ingests data from a comprehensive list of web sites ensuring good coverage of the main media sites for other geographies across all continents. This coverage will grow based on the needs, selecting the most suitable set of media sites in terms of languages, topics covered and geography.

Restrictions on content use and visibility within the platform and the SELMA consortium:

- Full content is only available for in-project users for testing purposes;
- Unlicensed content can be used, upon request, for project research according to copyright laws.

2.7 Provision of Technology-Specific Data

To develop specific technology components, the consortium has done both: annotate new data and collect existing data from its internal repositories. DW has provided data upon request when such data has been available (e.g., raw audio & video data, transcribed speech data for TTS, news articles and their summaries). As the technology components became available to end users for testing through the SELMA platforms, additional data has been gathered via user feedback. User feedback data is used primarily to improve entity linking and retrieval modelling.

To mitigate the risk of potential bias, discrimination and stigmatisation in the output models, we strived to make the training datasets as diverse and representative as possible by covering different topics, time periods, data sources (other news channels in addition to DW), languages, etc.

2.7.1 Raw Data

For training *wav2vec* speech recognition models for the selected SELMA languages, lists of DW audio/video recordings were collected by DW and provided to the SELMA technology partners. This was done via a private SELMA repository on GitHub, which contains only links to the actual audio/video data, which is publicly available from DW, YouTube and other websites.

For 22 languages, metadata of individual recordings is also gathered via the DW API. The metadata includes a text teaser for each recording, allowing using this data not only for training *wav2vec* models but also for modelling speech summarization.

Table 1 outlines the amount of DW audio/video data (with metadata, including text teasers) collected and provided for the development of language processing components of the SELMA platform (nearly 16k hours in total, more than 6M-word corpus of text teasers).

Language	Amount (hours)	Language	Amount (hours)
Amharic	47	Hindi	303
Arabic	1,094	Indonesian	232
Brazilian	2,920	Pashto	177
Chinese	425	Persian	487
Dari	133	Polish	148
English	435	Russian	1,314
French	752	Spanish	1,061
German	279	Turkish	1,385
Greek	121	Ukrainian	551
Hausa	6,284	Urdu	302

***Table 1** Amount of audio/video data provided by DW for technology development*

In addition to the audio/video data provided by DW, SELMA technology partners use additional data, for instance to process Pashto, Russian, Turkish and Dari (available for internal use only) for pre-training the large *wav2vec* models. However, for training speech summarization models, and for pre-training the SELMA multilingual open-source *wav2vec* model for 19 languages,[17] only the DW data was used.

Multilingual raw text data collected by the media-monitoring platform is used to train improved self-supervised text representations in the scope of WP2.

2.7.2 Transcribed Data

Specific datasets have been created by DW for training TTS models for selected DW languages: Brazilian Portuguese and Urdu. The latter is a low-resource language with respect to TTS – only four voices are accessible through Google's speech API.

The DW Brazil section has been producing two daily news bulletins since August 2020. Each bulletin is approximately 6 minutes long. In total, 870 audio bulletins have been collected, which results in approximately 87 hours of audio data. It should be noted that the Brazilian Portuguese dataset contains data from several speakers – DW Brazil news announcers – ranging

[17] https://huggingface.co/H2020SELMA

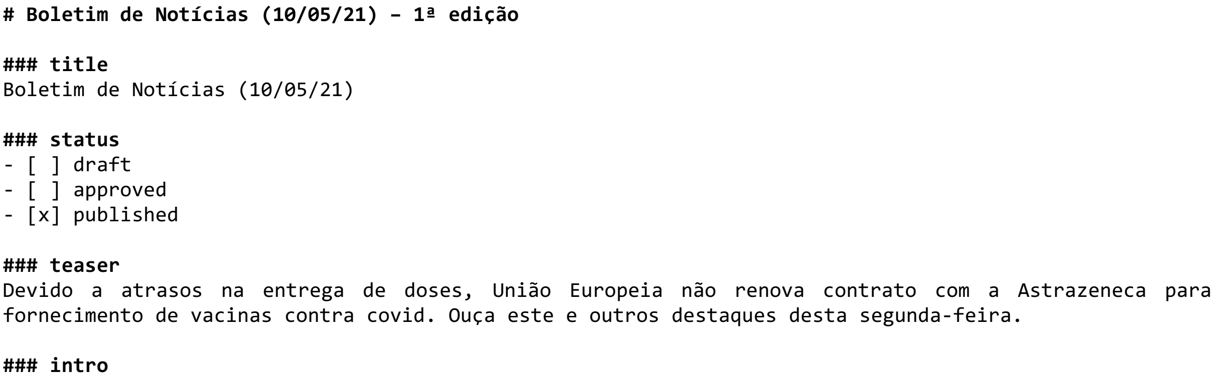
from 1 to 9 hours per speaker. As for the DW Urdu section, 10 hours of audio bulletins have been collected for a single speaker.

In addition to the audio files, there is an automatically generated subtitle file (SRT) available for each audio bulletin. For most of the bulletins, the script text written or edited by a DW journalist is also available. The scripts are provided in a markdown format, where the individual sections are separated by specific headers (see a sample in Table 2 and Figure 1).

No.	Header	Read out in the corresponding bulletin
1	Title	no
2	Teaser	no
3	Status	no
4	Intro	yes
5	Headlines	yes
6	Stories	yes
7	Sources	no
8	Outro	yes
9	Footnotes	no

***Table 2** Data fields provided for each DW Brazil news bulletin script*

Private SELMA project GitHub repositories are used to collect and manage the automatic subtitles and the manual transcripts for each language. A private LIA file server is used to store the audio data for further processing; however, the audio data is publicly available and ingested from DW's API that provides content to its mobile apps.



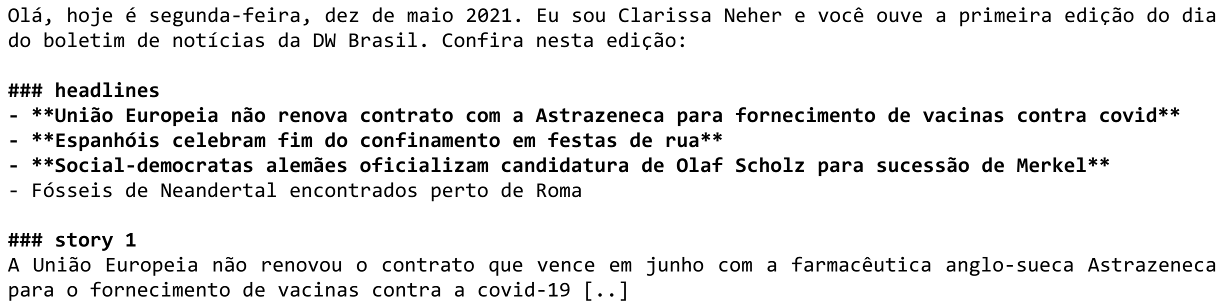
```
# Boletim de Notícias (10/05/21) – 1ª edição

### title
Boletim de Notícias (10/05/21)

### status
- [ ] draft
- [ ] approved
- [x] published

### teaser
Devido a atrasos na entrega de doses, União Europeia não renova contrato com a Astrazeneca para
fornecimento de vacinas contra covid. Ouça este e outros destaques desta segunda-feira.

### intro
```



```
Olá, hoje é segunda-feira, dez de maio 2021. Eu sou Clarissa Neher e você ouve a primeira edição do dia
do boletim de notícias da DW Brasil. Confira nesta edição:

### headlines
- **União Europeia não renova contrato com a Astrazeneca para fornecimento de vacinas contra covid**
- **Espanhóis celebram fim do confinamento em festas de rua**
- **Social-democratas alemães oficializam candidatura de Olaf Scholz para sucessão de Merkel**
- Fósseis de Neandertal encontrados perto de Roma

### story 1
A União Europeia não renovou o contrato que vence em junho com a farmacêutica anglo-sueca Astrazeneca
para o fornecimento de vacinas contra a covid-19 [..]
```

Figure 1 *Sample DW Brazil script in the markdown format*

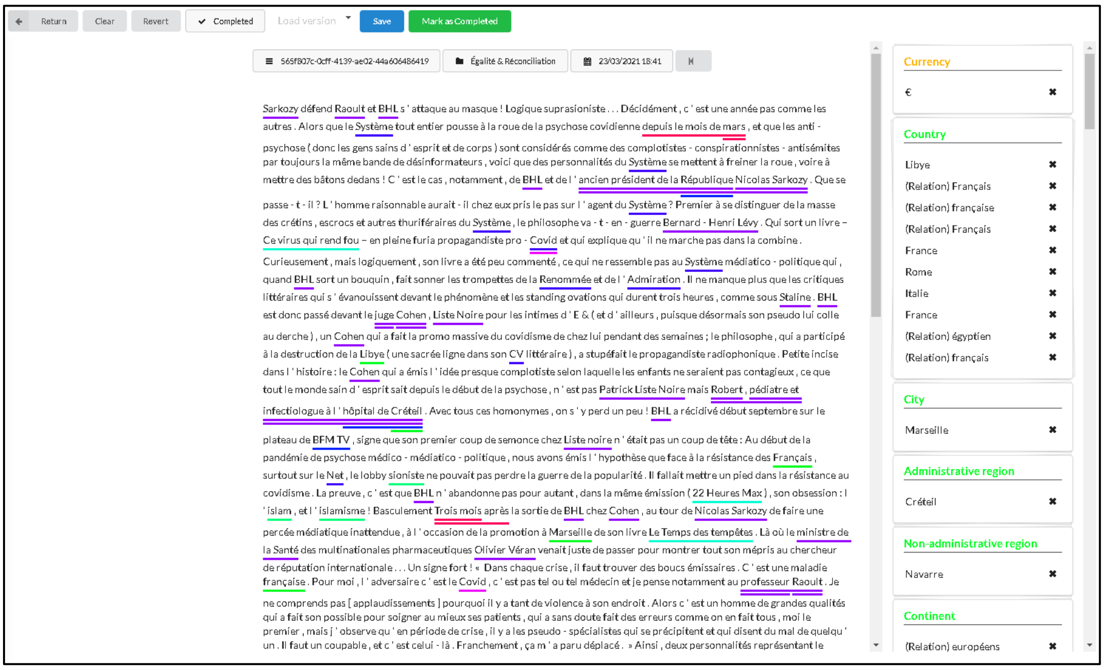
Additional datasets – annotated and corrected manuscripts with corresponding audio files in particular – are made available to the consortium partners (on request) by DW. This includes, for instance, a German dataset with single-speaker daily news reports, a collection of timecoded transcripts from audio or video in several languages (English, German, Russian, Hindi and Urdu), produced as corrected subtitles from DW productions. DW has also provided 10 hours of transcribed Amharic data and 5 hours of transcribed Bengali data.

2.7.3 Annotated Data

The NER-annotated data for the SELMA languages is provided to the consortium by DW, Priberam and IMCS, based on the specified requirements.

A mix of languages, both high-resourced and low-resourced, have been selected for NER annotation using the Priberam Annotation Tool (see Figure 2). This includes Ukrainian, Russian, Turkish, Dutch, Latvian. Language data annotated by Priberam (prior work): Portuguese, French, English, Spanish and German. Language data annotated by IMCS: Ukrainian (300 DW articles), Russian (160 DW articles) and Latvian (740 LETA articles).[18] Language data annotated by DW: Turkish (100 DW articles) and Dutch (50 DW articles).

[18] IMCS has received permission form the Latvian news agency LETA to use its articles in the NER-annotated Latvian dataset and to distribute the dataset to the research community.



Sarkozy défend Raoult et BHL s ' attaque au masque ! Logique suprasioniste . . . Décidément , c ' est une année pas comme les autres . Alors que le Système tout entier pousse à la roue de la psychose covidienne depuis le mois de mars , et que les anti - psychose (donc les gens sains d ' esprit et de corps) sont considérés comme des complotistes - conspirationnistes - antisémites par toujours la même bande de désinformateurs , voici que des personnalités du Système se mettent à freiner la roue , voire à mettre des bâtons dedans ! C ' est le cas , notamment , de BHL et de l ' ancien président de la République Nicolas Sarkozy . Que se passe - t - il ? L ' homme raisonnable aurait - il chez eux pris le pas sur l ' agent du Système ? Premier à se distinguer de la masse des crétins , escrocs et autres thuriféraires du Système , le philosophe va - t - en - guerre Bernard - Henri Lévy . Qui sort un livre – Ce virus qui rend fou – en pleine furia propagandiste pro - Covid et qui explique qu ' il ne marche pas dans la combine . Curieusement , mais logiquement , son livre a été peu commenté , ce qui ne ressemble pas au Système médiatico - politique qui , quand BHL sort un bouquin , fait sonner les trompettes de la Renommée et de l ' Admiration . Il ne manque plus que les critiques littéraires qui s ' évanouissent devant le phénomène et les standing ovations qui durent trois heures , comme sous Staline . BHL est donc passé devant le juge Cohen , Liste Noire pour les intimes d ' E & (et d ' ailleurs , puisque désormais son pseudo lui colle au derche) , un Cohen qui a fait la promo massive du covidisme de chez lui pendant des semaines ; le philosophe , qui a participé à la destruction de la Libye (une sacrée ligne dans son CV littéraire) , a stupéfait le propagandiste radiophonique . Petite incise dans l ' histoire : le Cohen qui a émis l ' idée presque complotiste selon laquelle les enfants ne seraient pas contagieux , ce que tout le monde sain d ' esprit sait depuis le début de la psychose , n ' est pas Patrick Liste Noire mais Robert , pédiatre et infectiologue à l ' hôpital de Créteil . Avec tous ces homonymes , on s ' y perd un peu ! BHL a récidivé début septembre sur le plateau de BFM TV , signe que son premier coup de semonce chez Liste noire n ' était pas un coup de tête : Au début de la pandémie de psychose médico - médiatico - politique , nous avons émis l ' hypothèse que face à la résistance des Français , surtout sur le Net , le lobby sioniste ne pouvait pas perdre la guerre de la popularité . Il fallait mettre un pied dans la résistance au covidisme . La preuve , c ' est que BHL n ' abandonne pas pour autant , dans la même émission (22 Heures Max) , son obsession : l ' islam , et l ' islamisme ! Basculement Trois mois après la sortie de BHL chez Cohen , au tour de Nicolas Sarkozy de faire une percée médiatique inattendue , à l ' occasion de la promotion à Marseille de son livre Le Temps des tempêtes . Là où le ministre de la Santé des multinationales pharmaceutiques Olivier Véran venait juste de passer pour montrer tout son mépris au chercheur de réputation internationale . . . Un signe fort ! « Dans chaque crise , il faut trouver des boucs émissaires . C ' est une maladie française . Pour moi , l ' adversaire c ' est le Covid , c ' est pas tel ou tel médecin et je pense notamment au professeur Raoult . Je ne comprends pas [applaudissements] pourquoi il y a tant de violence à son endroit . Alors c ' est un homme de grandes qualités qui a fait son possible pour soigner au mieux ses patients , qui a sans doute fait des erreurs comme on en fait tous , moi le premier , mais j ' observe qu ' en période de crise , il y a les pseudo - spécialistes qui se précipitent et qui disent du mal de quelqu ' un . Il faut un coupable , et c ' est celui - là . Franchement , ça m ' a paru déplacé . » Ainsi , deux personnalités représentant le

Currency

€ ✖

Country

Libye ✖
(Relation) Français ✖
(Relation) française ✖
(Relation) Français ✖
France ✖
Rome ✖
Italie ✖
France ✖
(Relation) égyptien ✖
(Relation) français ✖

City

Marseille ✖

Administrative region

Créteil ✖

Non-administrative region

Navarre ✖

Continent

(Relation) européens ✖

Figure 2 Sample NER-annotated data

The highly inflected and topologically different Latvian was the first additional language (to the Priberam's prior work) for which a NER-annotated dataset was created within SELMA. As our experience shows, after manual annotation of nearly 750 articles using the fine-grained Priberam Named Entities Annotation Schema, it was possible to train an automatic NER tagger for Latvian with ~85% accuracy. We also observed that transfer learning would yield similar accuracy with even less training data (e.g. 100–300 articles per new language). This allowed IMCS to switch from manual annotation to manual curation in the case of Ukrainian and Russian languages, which significantly improved productivity. Moreover, since Ukrainian and Russian are typologically closely related languages, we achieved good results by first curating ~300 pre-annotated Ukrainian articles and then adding ~150 curated Russian articles. Similarly, a smaller dataset sufficed for DW annotation of Turkish and Dutch, using pre-annotated data, thus reducing the need for large volumes of data. To make the SELMA NER datasets diverse and representative, we used the DW API to select medium size articles of various topics from various time frames. The Latvian news articles were provided to IMCS by the Latvian information agency LETA.

2.8 User data

In Use Case 1 and Use Case 2 “User Personal Data” is collected according to the Terms of Service (ToS) of each platform in compliance with GDPR, which includes the “Data Indispensable for the Performance of the Contract”, where the client assumes responsibility to provide Priberam, ensuring in advance that it has the legitimacy to do so, namely by having the necessary authorizations from the owners of the “User Personal Data” (“Data Subjects” in the scope of GDPR).

In this context, the following definition apply:

- “Client”, means an individual or corporate entity that holds a valid subscription to access < platform> under an agreement.
- “Agreement” designates the agreement concluded by the Client’s acceptance of a commercial proposal for access to <platform>, or a formal contract entered between the Parties stipulating the terms and conditions for access to <platform>.
- “Data Indispensable for the Performance of the Contract”, means data provided by the Client and/or the Users to Priberam, namely personal data concerning the client and/or each user, as well as any other additional data necessary to safeguard the legitimate interests of the Parties, collected, recorded and managed in accordance with the provisions of Priberam's privacy policy.[19]
- “User Personal Data”, means data collected in the scope of “Data Indispensable for the Performance of the Contract”, namely the data needed to register and authenticate the user in the platform and the data generated during platform usage such as log data or customization data.

“User Personal Data” is collected, protected and stored with the following put in place:

a) it is stored within the European Union: in Priberam Servers hosted by AR Telecom [ART] in Portugal. ART is certificated under ISO 9001, ISO 14001, ISO 20000 and ISO

[19] Accessible at https://priberam.pt/Docs/Priberam_Politica_de_Privacidade.pdf

27001 (see https://www.artelecom.pt/certificacoes/), and only employees with production privileges at Priberam, which are bound by NDAs, can connect remotely into ART and have access to the data.

b) it is encrypted with measures implemented following the NIST guidelines[20] as a reference:

– Display name and email is encrypted at rest (database) with AES Symmetric Encryption, 32byte key (256bit). For generating the key, the system uses RFC2898 with SHA-512, a unique 16byte random salt per secret (distributed with secret) concatenated with a 16byte secret pepper unique to the application (not distributed with secret). The system uses a random initialization vector which is unique per secret).

– Passwords are never transferred or stored in plain text; they are hashed in the frontend with SHA-512 and unique known salt per user (email).

c) database backups may leave ART to ensure geographical diversity in the scope of the backups policy but will not leave EU and personal data goes encrypted there).

d) communication is done using encrypted HTTPS, and JWT is used for authentication.

Both platforms collect user feedback data to enable incorporation of models, based on the interactions of the users and the platforms. These interactions are not directly linked to user's personal data, but through an "account Id" – it is required to hold both the "account Id" and the database decryption key to fully establish the link between these interactions and the personal data. This data can be used to improve the results in certain tasks by automatic post editing. The improved models are used only on the scope of the user that provides the data, or if in the context of a multi-user contractual organization with the given consent of the performers.

Use Case 1 is collecting the following user interactions:

- Correction of NER spans and classification in news articles;
- Correction of linked entities;

[20] https://pages.nist.gov/800-63-3/sp800-63b.html

- Additional tags entered by the users for specific news articles with user defined taxonomies;
- Relevance given by users on the retrieval of news either by marking retrieved items as curated or rejected since unrelated.

Use Case 2 is collecting the following user interactions:

- User edits on transcriptions;
- User edits on translations;
- User edits on subtitles;
- User edits on voice over tasks.

3. Types of Generated Data

“Generation of data” in this report primarily refers to the production of data by the SELMA platform or any of its components:

- Speech transcripts of the multilingual broadcast content – generated by the ASR components.
- Synthesized speech for the multilingual broadcast content – generated by the TTS components.
- Machine-translated (MT) broadcast content (including ASR-generated transcripts) – generated by the neural MT and speech translation components.
- Named entity annotations, automatic summaries – generated by the named entity recognition/linking, abstractive text/speech summarization components.
- Clustering and storyline detection on news articles.
- News article classification with IPTC subject codes.

We distinguish between the following categories of data that is generated within the project:

- Content data generated during media monitoring (UC1) and news production (UC2), as well as testing of the SELMA platform (UC0). This is typical broadcast data that remains copyright protected. See Figure 1 (Section 2.7.2) for an example.
- Specific output formats with regard to particular steps in the SELMA language processing pipelines. This includes transcriptions, translations, summaries, annotations, statistical data, and usually includes broadcast content as well. See Figure 3 (Section 4.1) for an example.
- Software, acoustic and language models, task specific models, lexicons and ontologies, linguistic annotations and user feedback. See Figure 2 (Section 2.7.3) for an example.
- Academic research publications (journal articles, conference papers, preprints).

See Section 6 for complementary details regarding sharing of generated data.

4. Data and Metadata Standards

This section briefly describes standards and formats used in the project for handling, referencing and interchanging data within the SELMA platform and for robust and scalable automatic ingestion of news items into the platform from DW and other sources.

4.1 Data Identifiers and Internal Data Format

All data units stored in the SELMA monitoring platform (news and media items, both original and derived content; semantic annotations, like named entity mentions; etc.) are identified by universally / globally unique identifiers (UUID / GUID). These identifiers are generated and assigned by the platform upon data ingestion (to the source content) and during data processing (to the derived or enriched content).

The SELMA platform internally uses a JSON data structure (see a simplified illustration Figure 3), agreed between the consortium partners, which encodes references to source content and contains the output content automatically generated by SELMA language processing components (workers).

```
{
  "workflowId": "f3bd989f-bbdb-4851-857c-549b884e3641",
  "jobNodes": [ {
    "id": "abba189f-bbdb-4851-857c-549b884e3641",
    "jobData": {
      "Worker": "ASR-LV",
      "Text": "selma.ailab.lv:2020/files/4963f238-9b83-4b37-9553-dc8ae608d719"
    },
    "jobResult": {
      "words": [
        { "word": "no", "confidence": 1.000, "time": 1.039, "duration": 0.169 },
        { "word": "darba", "confidence": 1.000, "time": 1.209, "duration": 0.309 },
        { "word": "uz", "confidence": 1.000, "time": 1.519, "duration": 0.079 },
        { "word": "mājām", "confidence": 0.823, "time": 1.599, "duration": 0.489 },
        ...
      ]
    }
  },
  {
    "id": "abba289f-bbdb-4851-857c-549b884e3641",
    "dependencies": [ "abba189f-bbdb-4851-857c-549b884e3641" ],
    "jobData": { "Worker": "ASR-Punctuation" },
    "jobResult": { "text": "No darba uz mājām mēs braucām vienā un laikā visu gadu. " }
  },
  {
    "id": "abba489f-bbdb-4851-857c-549b884e3641",
    "dependencies": [ "abba289f-bbdb-4851-857c-549b884e3641" ],
    "jobData": { "Worker": "EasyNMT", "source_lang": "lv", "target_lang": "de" },
    "jobResult": {
      "alignment": [ {
        "text": "No darba uz mājām mēs braucām vienā un tai pašā laikā visu gadu.",
        "translation": "Wir fuhren das ganze Jahr über zur gleichen Zeit von [..]."
      } ]
    }
  } ]
}
```

Figure 3 *A JSON data snippet illustrating the SELMA internal data exchange format*

The JSON data format and the internal data flows are further detailed in D4.1 "Platform architecture and API documentation".

4.2 Text Feeds

The most common format to distribute news content is syndication via RSS and ATOM feeds. DW articles are available via RSS, ready for ingestion into the SELMA platform.

An alternative method to disseminate news content is the use of XML sitemaps or news sitemaps. This also applies to DW content.

As RSS, ATOM and XML sitemaps are standardized formats used by many publishers, they represent the preferred method to ingest content into the platform.

Alternatively, we can access DW's content through its proprietary API. This is a custom method that cannot be easily transferred to other news providers and is therefore considered being a last-resort fallback, in case the methods described above are inadequate, or insufficient to collect the full content of a news item.

As a last resort, news links are gathered by scraping news links from specific web sites using a rule-based (pattern-matching) system to collect relevant pages.

4.3 Audio & Video Feeds

Just as with the distribution of article texts, a common way to syndicate audio and video content is the use of podcast feeds which in turn use the RSS format as described above.

Much of DW's content as well as content provided by other news sources is accessible via podcast feeds. For relevant DW content that is not published as podcast feed, the DW API is used as fallback.

4.4 Entity Identifiers and Properties

For named entity linking (based on the named entity recognition output), we use the widely acknowledged open Wikidata knowledge graph and its entity identifiers (e.g. Q3874799 for Volodymyr Zelenskyy[21]).

Following our decision (see Section 2.6) not to collect and analyze any sensitive personal data categories, for each entity representing a person, we collect only the following properties from Wikidata: date of birth (P569), binary sex or gender (P21), continent (P30; to support analysis by this property only at the Global South vs. Global North level).

[21] https://www.wikidata.org/wiki/Q3874799

The extracted entity metadata is stored at the document (a news item) level in the MONITIO platform (Use Case 1). A set of entities with their properties is linked to where these entities are mentioned in the text.

5. Data Storage, Preservation, Reuse and Sharing

Media monitoring data (text, audio and video, metadata) produced by DW and collected by Priberam (from external sources) is directly and automatically ingested into the SELMA platform (MONITIO) repositories for development, testing and demonstration purposes. Additionally, DW provides access to its APIs to the technical partners for automatic retrieval of DW's multilingual content in case of specific data ingestion scenarios (e.g., to collect text data for named entity annotation).

Technology-specific data (text, audio and video, annotations) produced and collected by DW, Priberam and IMCS is stored in private SELMA GitHub repositories managed by DW and used by all consortium partners. It contains selected broadcast content for developing and testing the language processing components of the SELMA platforms and tools:

- *SELMA-project/brasil-noticias-scripts* and *SELMA-project/dw-urdu-news* – contains scripts of news bulletins produced by DW Brazil and DW Urdu, together with generated subtitles and with links to the respective audio/video files that are publicly available from DW and YouTube websites, cannot be released as open data (see also Section 2.5.2 and 2.7.2 for more details).
- *SELMA-project/DW-AV-Data* – lists of DW audio/video recordings, i.e., lists of links to the actual audio/video data, which is publicly available from DW, YouTube and other websites, cannot be released as open data (see also Section 2.5.1 and 2.7.1).
- *SELMA-project/youtube-audio-data* – additional lists of audio/video data, publicly available from YouTube, cannot be released as open data (see also Section 2.5.1 and 2.7.1).
- *SELMA-project/HNNER_Torch* – NER-annotated datasets created within the project; has been partially released (the Latvian dataset) for the research community via the European language resource repository CLARIN[22] (see also Section 2.5.3 and 2.7.3).

[22] http://hdl.handle.net/20.500.12574/98

Other SELMA data repositories:

- Datasets created for training NER and NEL models are primarily stored in a private database maintained by Priberam; some of these datasets are a prior work by Priberam and will not be released in the public domain, others created by DW are considered for release but pending a license approval; the Latvian dataset created by IMCS is released for the research community (approved by the IPR holder LETA).
- Ingested and annotated data on UC1 (MONITIO) are stored in a private Postgres database at Priberam premises. This data will not be released outside of the consortium, unless proper agreements with the original content producers are in place for the intended use.
- All data about user feedback respects the policies defined in Section 2.8. If the data is needed for research purposes it is exported from the platforms losing all connections the original users, by using another layer of anonymization where there is no connection between the generated identifiers and the platform internal identifiers.

The technical partners have used selected DW datasets (described in Section 2.7) for specific training and testing of language models and other language processing components used for ASR, TTS, NER and summarization. For these activities, the necessary datasets have been retrieved from the DW repositories and stored on the partner servers.

The technical partners retrieved the technology-specific data from the shared repositories and used it for development and testing purposes, while the SELMA platform itself (MONITIO) ingests monitoring data via content feeds and APIs, after which the data is stored on the platform's servers. Production instances of the SELMA platforms (UC1 and UC2) are managed by Priberam, and, consequently, ingested content is stored on its servers. Technical partners (IMCS and Priberam) have set up a development environment for DW to test Use Case 0, Use Case 1 and Use Case 2 applications and their components. Ingested UC1 and UC2 content is stored in a database for further processing (UC0 content is not stored). Downstream tasks performed on the data enrich the data and store the information together with the original documents. When the required tasks are applied, the data is indexed and made available for the front-end. It should be noted that any cloud services that have been selected for the use in SELMA (e.g. Azure) are always located in the EU to respect GDPR.

To ensure sustainability and reuse of the SELMA open datasets and models, and to facilitate their discovery by the research community, the SELMA Foundational Multilingual Model for ASR (see Section 2.7.1) is made available via the public SELMA Hugging Face repository,[23] while the SELMA Latvian NER Dataset (see Section 2.7.3) is shared for academic use (due to IPR protection; requires eduGAIN authentication to download the dataset) via the European language resource and technology research infrastructure CLARIN.[24] Similarly, the SELMA Open-Source Platform is also registered at CLARIN[25] (in addition to its public GitHub repository[26]) and is discoverable via the pan-European federated search facility – CLARIN Virtual Language Observatory.[27]

[23] https://huggingface.co/H2020SELMA

[24] http://hdl.handle.net/20.500.12574/98

[25] http://hdl.handle.net/20.500.12574/97

[26] https://github.com/SELMA-project/UC0-OpenSource

[27] https://vlo.clarin.eu

6. Policies for Data Access and Sharing

There are different kinds of categories of data that has been collected or generated during the project, with different levels and conditions for access and sharing:

- Original broadcast data is copyright-protected and as stipulated in the Consortium Agreement, is provided by DW only for use by the consortium partners for the duration of the project. It can therefore not be shared outside the consortium or after the project. Some demo material has been selected for public viewing in agreement with DW.
- Data generated during media monitoring is typically owned by the broadcaster; therefore, the consortium does not have the right to share this as open research data. However, negotiations with DW are pending approval to release the DW NER-annotated datasets for the wider research community. The Latvian NER dataset has already been released with the approval by the Latvian news agency LETA.
- Specific output formats following a particular step in the SELMA language processing chain are open as such, however, the output data itself usually includes (or is derived from) broadcast content and therefore cannot be shared as open data. This includes automatic transcriptions, translations, summaries, annotations and statistical data.
- Specific software and language models, i.e., the SELMA Open-Source Platform and the SELMA Foundational Multilingual Wav2Vec2.0 Model are made available in the public domain.
- Academic research publications are made available as open-access via institutional or public repositories and via the OpenAire system.

Only publicly available news items and published media content are targeted for data gathering. Ingestion of social media data is restricted to news and published media by public broadcasters for which we have reached a license agreement. The platform does not ingest any data from social media users (comments, replies, etc.): all efforts are made to avoid collecting user comments or other user-generated personal data. For instance, some news items published on a broadcaster website may contain embedded tweets; the SELMA web-scraper algorithms try to detect and remove such embeddings from the collected news content. Only data necessary for the completion of the project has been stored.

Data security procedures were established for each partner dealing with SELMA datasets. Regarding the non-sensitive Wikidata properties of named entities, such as binary gender and age, this data is collected and stored to support the SELMA diversity use case (cf. D1.1), while mitigation procedures of potential ethical issues are described in D8.1.

Access to the SELMA data repositories and to the SELMA UC1 and UC2 platforms (populated with data) is secured using SSL via the HTTPS protocol and requires authentication (except for the public SELMA open-source platform).

Each of the partners acts as a data processor for the data needed for their own activities. Each partner acts as a data controller of their own data and their employes. Priberam is the data controller for UC1 and UC2 regarding the user data needed for the operation of the platforms. Priberam is the data processor for ingested data extracted from the media publishers web sites and the data controller for the metadata automatically produced to enrich the ingested data.

Besides the guidelines outlined in this document, all consortium partners dealing with data, including provision, use, processing and storing, comply with data protection regulations for their organization and country. Partners have been responsible for seeking advice from their respective local data protection authorities.

See D8.1 “Ethics Deliverable” for more details on SELMA measures to ensure privacy and personal data protection.

7. Conclusion

The final Data Management Plan describes the SELMA data management strategy and actions, as discussed and agreed by the consortium partners. It addresses identified issues and aspects related to the collection, generation and sharing of data, including intellectual property rights and personal data protection, as well as long-term preservation and re-use.

To facilitate data reuse and thus ensure its sustainability, specific software, language models and datasets developed in the project are made available for public or research use, when and if it has not been in breach of copyright and personal data protection.

The are no major changes in the final DMP compared to the interim DMP (D6.3). It provides updated and more elaborate information, as well as the final decisions made regarding availability of the open-source software, language models and datasets.