Research and Innovation Action (RIA) H2020 – 957017

SELMA

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D6.1 Initial Data Management Plan

Work Package	6
Responsible Partner	IMCS
Author(s)	Normunds Grūzītis (IMCS), Guntis Bārzdiņš (IMCS)
Contributors	Andreas Giefer (DW), Afonso Mendes (Priberam), Yannick Estève (LIA), Kay Macquarrie (DW), Peggy van der Kreeft (DW)
Version	1.0
Contractual Date	30 June 2021
Delivery Date	30 June 2021
Dissemination Level	Public

Version History

Version	Date	Description
0.1	29.04.2021	Initial Executive Summary, Table of Contents, Introduction
0.2	27.05.2021	First draft with template for contributions
0.3	10.06.2021	Updated and extended draft
0.4	14.06.2021	Full draft ready for review
0.5	22.06.2021	Integrated contributions from all partners
1.0	29.06.2021	Final version for submission

Executive Summary

The Data Management Plan provides an analysis of the main elements of the data management policy that will be used by the SELMA consortium with regard to all 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, and 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) to be able to handle up to ten million news items per day. The media monitoring platform will be 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 will also be applied to a **News Production Tool** (Use Case 2), using enriched models for transcription (ASR) and translation (MT), giving journalists in an operational editorial environment a multilingual tool that will be able to learn over time.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

The Data Management Plan functions as a central tool for risk mitigation associated with data protection. The initial Data Management Plan includes the following aspects:

- A description of what research and innovation activities of the project will use which data, and a description of who will be responsible for handling, storing, and destroying the data (data processing).
- A description of the purpose of SELMA research and innovation, to make clear that there is a substantial public interest in the work of the project.
- A description of the safeguards that we will put in place.
- Identification of the countries in which data will be processed or will reside, together with an understanding of the national privacy and data protection regulations, and engagement with the relevant data protection agencies.

The Data Management Plan will also integrate outcomes of the privacy impact assessment (see D8.1 Ethics Deliverable): information flows in the project, identification of the privacy and related risks, actions taken by SELMA to reduce the identified risks.

This is the initial version of the Data Management Plan. This document will be updated within the course of the project's development. There will be two further iterations (D6.3 and D6.5) which will elaborate on the issues covered. The issues addressed here are also part of the ethics, project management and evaluation reports.

2. Types of Data Collected

SELMA develops an open-source platform for dealing with large volumes of data across many languages and different media types. It has a range of technologies that are implemented, including automated speech recognition and synthesis, machine translation, named entity recognition and linking, and summarization.

Data is being 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, Indonesian, Kiswahili, Macedonian, Pashto, Persian, Polish, Portuguese for Africa, Portuguese for Brazil, Romanian, Russian, Serbian, Spanish, Turkish, Ukranian, Urdu.

The project consortium includes two 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.

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

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

The first use case is targeted by both DW and Priberam, while the second use case is targeted solely by DW.

“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 SELMA language processing components.

2.1 Data Types

Data for SELMA is being collected at several levels:

- By the intended use: ingestion data, training data, test data.
- By the language processing technology: speech recognition, speech synthesis, machine translation, named entity recognition and linking, summarization.
- By data type: metadata, text, audio & video.
- By delivery type: batch data (incl. text streams); no audio or video live streams.
- By language: the 30+ SELMA languages.
- By content and language data provider/user: DW, Priberam, other partners.
- By user feedback (e.g., through editors while correcting transcripts and through platform usage).

We divide data requirements and data provision into two target groups: the regular content for media monitoring, and the specific datasets for the development of language technology components, namely, speech recognition and synthesis, contextual representations, classification, storyline clustering, retrieval and named entity recognition and linking.

2.2 Requirements for Monitoring Data

Since SELMA Use Case 1 deals with data monitoring, such data is essential for the prototype development and assessment, user validation and scalability testing.

Different types of data are involved, but only one type of delivery is targeted within the SELMA project:

- Types of data: metadata, text data, audio & video data.
- Types of delivery: batch data (i.e., audio & video live streams are not involved).

These aspects and challenges are detailed further in this report.

2.3 Requirements for Technology-Specific Data

Requirements and specifications for technology-specific datasets are being gathered within WP2 and WP3, detailing what type of data is needed, and how much. The SELMA partners are directly supporting 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.

All the 30+ languages will eventually be supported by the SELMA platform, either by in-house development of the respective language processing components or by exploiting third-party APIs. The focus during the initial stage of the project is on a selected mix of high- and low-resourced languages: English, German, French, Portuguese (both versions), Spanish, Turkish, Polish, Indonesian, Chinese (both versions), Hindi, Persian, Arabic, Greek, Pashto and Latvian.

2.3.1 Transcribed Data

For the automatic speech recognition and transcription (ASR), already existing datasets (transcribed speech corpora, see below) and APIs will be used within the SELMA project, and no additional datasets will be collected for the ASR development.

For the automatic text-to-speech synthesis (TTS), however, speech datasets of a limited amount will be collected for some languages (e.g., Brazilian Portuguese) for which appropriate existing datasets or APIs 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[1], CSS10[2] and CommonVoice[3] speech datasets, the TED-LIUM3 speech dataset[4], and the very recent Spotify Podcast Dataset[5] will be considered;

ii. restricted-access datasets like QUAERO and ETAPE are available for internal use;

b) the 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) the current work on transfer learning of acoustic and language models for targeting selected 3 of the low-resourced priority SELMA languages;

d) the creation of limited amounts of speech datasets for selected SELMA languages; at least 20-30 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 will use for this purpose editorially corrected manuscripts/transcripts from its editorial HLT platform.

For each audio file in a speech corpus, a correct transcription of the spoken text is required. A fallback solution is subtitled data, i.e., loose transcription that has to be provided if the exact transcript is not available and it would be too labor-intensive to provide it.

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 is UTF-8. The specific data format for each language has to be clarified between the data provider and the technical partners.

---

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

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

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

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

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

2.3.2 Annotated Data

For named entity recognition (NER) and linking (NEL), creation of a multilingual dataset is considered by the SELMA consortium. For a selected subset of SELMA languages for which compliant prior datasets are not available, a representative set of approximately 3000 documents per language could be manually annotated by SELMA partners according to a common NER and NEL annotation schema. The amount of annotated documents could be reduced by using the language transfer mechanisms researched within the project.

Regarding the document selection for each language, the focus will be on news items and bulletins, i.e., broadcast news which 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 random public persons (mentions of person names and related entities).

As for the common annotation schema, the Priberam Named Entities Annotation Guidelines (see Annex) is considered as the fundament and orientation for the SELMA dataset.

The SELMA multilingual dataset for training and evaluation of NER systems would have a significant impact on the NLP research community, if it is released as open data by the SELMA consortium. NER-annotated data created within the project will be shared with an open license, unless prohibited by copyright restrictions. This would exclude the prior data annotated for Portuguese, French, English, Spanish and German that will be used in the scope of the project but will not be released with an open license.

Since SELMA partners already use private GitHub repositories for development purposes, the sharing of the NER dataset will be done via a public GitHub repository which then has to be disseminated to reach a wider research community. Additionally, the dataset could be distributed via the European research and innovation infrastructures for language resources and

technology: CLARIN[6], European Language Grid[7], and others. This would ensure not only sustainability of the essential language resource, but would also facilitate its discovery, use and citation (with a persistent identifier).

2.4 Provision of Monitoring Data

Batch data will be periodically collected and ingested into the SELMA platform for the media monitoring use case (Use Case 1). Audio and video that will be collected through the ingestion pipeline but no live stream data will be ingested since it is too noisy. The platform will be tested with up to 10 million news items per day.

First, the monitoring data will be provided by DW, covering all the 30+ SELMA languages. There are three alternatives for the automatic DW content ingestion into the SELMA platform:

- By ingesting RSS feeds;
- By making API calls;
- By crawling XML site maps,
- By scraping document links from specific internet sites with none of the above possibilities.

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

Second, and most important, news items from other media sites will be collected and provided by Priberam – by scraping news portal content based on XML site maps, by ingesting RSS, news sitemaps, sitemaps and by scrapping links from specific sites. Since media publishers are increasingly publishing unique content on social media platforms like Twitter, Facebook,

[6] https://www.clarin.eu/content/services

[7] https://live.european-language-grid.eu/

Instagram, TikTok and YouTube, we are applying for access to gather data from public media pages on those platforms. We will not collect personal data from social media users except as aggregated data that will be used to quantify the reach of particular media items or media producers.

DW content will ensure testing the multilingual aspects of the SELMA platform, but it will not be sufficiently big data 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 will be collected and ingested into the SELMA platform.

The system will try to cover as much of the published media as possible. Currently we are already ingesting more than 5000 target sites covering almost completely Portugal and Spain and the main media sites for other geographies in Europe, Latin America and Oceania. The coverage will grow based on the needs of the project selecting the most suitable media sites according to the needs of the project in terms of languages, topics covered and geography.

2.5 Provision of Technology-Specific Data

In order to develop specific technology components, the consortium will either annotate new data or collect existing data from its internal repositories. DW will provide data upon request when such data is available, (e.g., audio & video transcription data, speech data for TTS, news summaries). The consortium will continue to annotate data for NER purposes, extending the number of languages already available. As the technology components became available for end-users through the SELMA platform, additional data will be gathered via user feedback. User feedback data will be used at least for entity linking and retrieval modelling.

2.5.1 Transcribed Data

One specific dataset has been identified so far for speech data, i.e., a Brazilian Portuguese transcription dataset, collected by DW.

The DW Brazil section has been producing two daily news bulletins since August 2020. Each bulletin is approximately 6 minutes long. As of 10 June 2021, 410 bulletins have been collected, which results in approximately 2460 minutes (41 hours) of audio data. It should be noted that

the current Brazilian Portuguese dataset contains data from several speakers – DW Brazil news announcers. However, one of the speakers is dominant and will be filtered out for the TTS needs, therefore the actual size of the dataset will be clarified later.

In addition to the audio files, there is an automatically generated subtitle file (SRT) available for each bulletin. For about half of the bulletins (238 at the time of writing), the original script file written by the DW journalists is also available. More scripts are being added. The scripts are provided in a markdown format, where the individual sections of the bulletins are separated by markdown headers (see Table 1 and Figure 1).

Table 1 Data fields provided for each DW news bulletin script

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

A private SELMA project repository on GitHub is used to collect and manage the automatic subtitles and the manual transcripts. A private LIA file server is used to store the audio data. Note that both the audio data and the automatic subtitles are available from DW Brazil's YouTube channel.

```
# 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 A sample script in the markdown format

Other datasets, in particular annotated and corrected manuscripts with corresponding audio files, will be made available to the consortium by DW. This includes a German dataset with single-speaker daily news reports. Also, a collection of timecoded transcripts from audio or video in several languages, including English, German, Russian, Hindi and Urdu, produced as corrected subtitles from DW productions, will be provided.

2.5.2 Annotated Data

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

At the first stage, a mix of 10 priority languages, both high-resourced and low-resourced, have been selected for DW annotation: Arabic, Chinese (both versions), Greek, Hindi, Indonesian, Pashto, Persian, Polish, Turkish. Language data annotated by Priberam (prior work; see Figure 2): Portuguese, French, English, Spanish and German. Language data to be annotated by IMCS: Russian, Ukrainian and Latvian.

Priberam already annotated data for Portuguese, German, French, Spanish and English. DW is annotating for Arabic and IMCS for Latvian. More languages will be gradually annotated and added to the dataset.

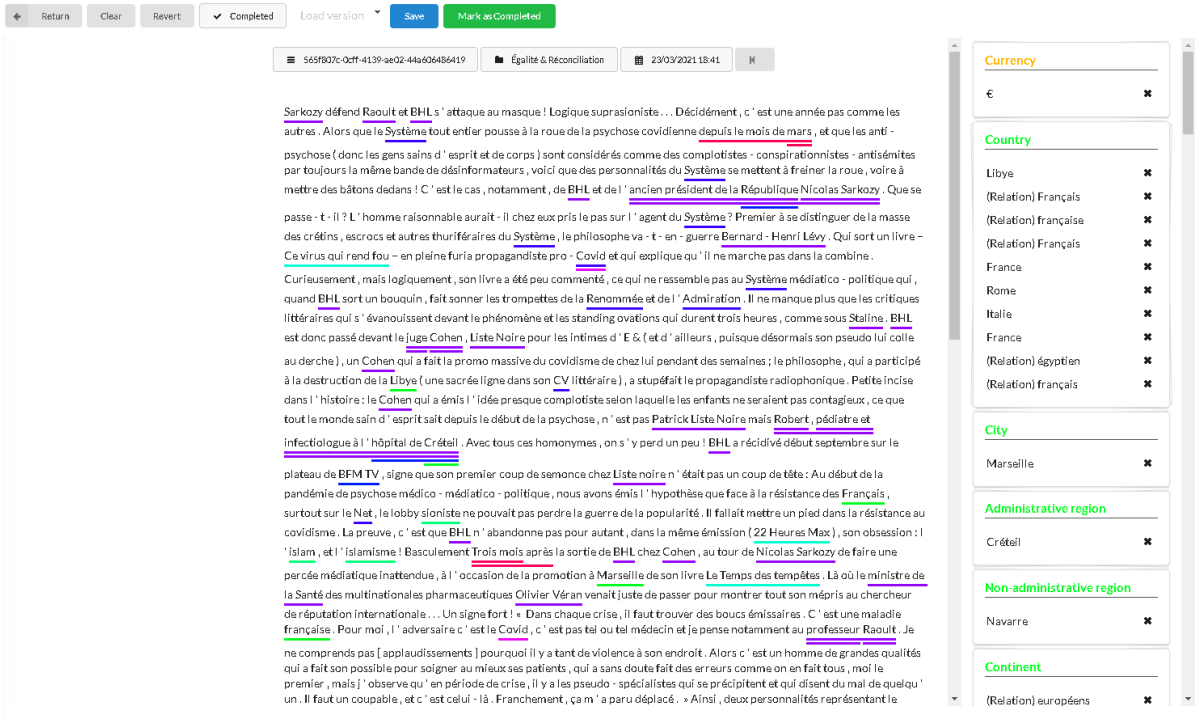

Figure 2 Sample NER-annotated data

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 component.
- Named entity annotations, automatic summaries, etc. – generated by the semantic parsing and summarization components.

We distinguish between the following categories of data that will be generated during the project:

- Content data generated during media monitoring (Use Case 1) and news production (Use Case 2). This is typical broadcast data that remains copyright-protected. See Figure 1 (Section 2.5.1) for an example.
- Specific output formats with regard to particular steps in the SELMA language processing pipeline. 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 2.5.2) 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 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 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 will be further detailed in D4.1 “Platform architecture and API documentation”.

4.2 Text Feeds

The most common format to distribute news content is the syndication via RSS and ATOM feeds. DW is making its articles 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 or other news 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 that the methods described above are inadequate, or insufficient to collect the full content of a news item.

As a last resort, news links will be 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 can be used as fallback.

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) will be directly and automatically ingested into the SELMA platform repositories for development, testing and demonstration purposes. This data will be accessible to all consortium partners. Additionally, DW will provide access to its APIs to the technical partners for automatic retrieval of DW's multilingual content in case of specific data ingestion scenarios.

Technology-specific data (text, audio and video, annotations) produced and collected by DW, Priberam and IMCS will be stored in a SELMA GitHub repository managed by DW and used by all consortium partners. It will contain selected broadcast content for developing and testing the language processing components of the SELMA platform.

The technical partners will use selected datasets (like the Brazilian Portuguese dataset described in Section 2.5.1) for specific training and testing of language models and language technology components, e.g. for ASR, punctuation, MT, NER and summarization. For these activities, the necessary datasets will be retrieved from the DW repository and stored on the partner servers.

The technical partners will retrieve the technology-specific data from the shared repository and will use it for development and testing purposes, while the SELMA platform itself will ingest monitoring data via content feeds and APIs, after which the data will be stored on SELMA platform servers, initially maintained by IMCS. Production instances of the SELMA platform will be managed by DW and Priberam, and, consequently, ingested content will be stored on their servers. Technical partners (IMCS and Priberam) will set up a development environment for DW to test Use Case 1 and Use Case 2 applications. Full-scale deployment at DW is expected to require hosted computing resources from AWS, Azure or similar cloud services. Ingested content is stored in a database for further processing. 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 frontend.

Data preservation and sharing options after the project will be discussed in the final version of Data Management Plan, when the final output form of data will become visible. However, it is

envisaged that the SELMA platform is usable and customizable after the project, as it is especially geared towards Use Case 1 and Use Case 2.

To ensure data sustainability and reuse, and to facilitate its discovery, selected datasets (like the multilingual NER-annotated dataset) created within the project and useful to the larger research community will be considered for sharing also via the European research and innovation infrastructures for language resources and technology, like CLARIN and ELG.

The data produced during the course of the project will be available in accordance with the Consortium Agreement and license agreements. Data reuse and sharing will be ensured as much as possible and will primarily apply to software, certain lexicons and corpora.

6. Policies for Data Access and Sharing

There are different kinds of categories of data that will be 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 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 will be 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 rights to share this as open research data. However, negotiations will be opened with DW with the aim of releasing particular data sets for specific research use.
- 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 (e.g. aggregation of social media reach).
- Software, language models, lexicons, linguistic annotations (like the named entity annotations illustrated in Figure 2) and other technology-specific training datasets, etc. data will be made available as open as possible. We shall endeavor to publish and make open access derived data when this is not in breach of copyright.
- Academic research publications will be made available as open access via institutional repositories and via the OpenAire system.

Regarding measurements for the protection of personal data, the SELMA project and technology platform does not focus on acquiring and processing personal data. However, broadcast content may contain some data that can refer to random individuals. SELMA will apply standard methods for the protection of such personal data, in particular regarding the gathering, storage, retention and the destruction of (personal and other) data.

Only publicly available news items and published media content will be targeted for data gathering. Ingestion of social media data will be restricted to news and published media and its numerical reach data. All efforts will be 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 to the completion of the project will be stored.

Security procedures will be established for each partner dealing with data. Access to the SELMA data repositories and to the SELMA platform (populated with data) will be secured using SSL via the HTTPS protocol and will require authentication.

It is understood that the consortium as a whole are joint data controllers in this project. All consortium partners dealing with data, including provision, use, processing and storing, will make their best efforts to comply with data protection regulations for their organization and country. Partners are responsible for seeking advice from their respective local data protection authorities.

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

7. Conclusion

The initial Data Management Plan (D6.1) provides the basis for the SELMA project data management strategy and planning, as discussed and agreed by the consortium partners. It addresses so far identified issues related to the collection and generation of data, data set identifiers and descriptions, standards, data sharing, property rights and personal data protection, as well as long-term preservation and re-use.

To facilitate data reuse and, thus, ensure its sustainability, software, language models, and derived technology-specific training datasets developed within the project will be made available as open and accessible to the research community as possible when this is not in breach of copyright and personal data protection.

This is the first of three iterations; the interim update is due at the project midterm (M18) and the final version is due at the end of the project (M36). Data collection, generation and processing are key areas in the SELMA project and will be discussed, elaborated and further specified throughout the project.

8. Annex – Named Entity Annotation Guideline

priberam

priberam.com • flip.pt • legix.pt

Priberam Named Entities Annotation Guidelines

Digitally signed by Priberam Informatica, S.A.
DN: c=PT, o=Priberam Informatica, S.A., 2.5.4.97=VATPT-502237740, ou=Certificate Profile - Organization, ou=Authentication - Identity Authenticated, email=veritas@priberam.pt, cn=Priberam Informatica, S.A.
Date: 2021.03.22 12:04:27 Z

Priberam Informática, S.A.
Registada na C.R.C. de Lisboa sob o NIPC
Capital Social: 60.000 Euros | NIPC (VAT ID): PT 502 237 740

Sede | HQ: Alameda D. Afonso Henriques, 41 – 2.º
1000-123 LISBOA
PORTUGAL

+351 217 817 260
info@priberam.com