



SELMA

Stream Learning for Multilingual Knowledge Transfer

<https://selma-project.eu/>

D6.3 Interim 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	3.0
Contractual Date	30 June 2022
Delivery Date	29 June 2022, 18 November 2022 , 31 March 2023
Dissemination Level	Public

Version History

Version	Date	Description
0.1	06.06.2022	First D6.3 draft
0.2	16.06.2022	Updated D6.3 draft
0.3	24.06.2022	Integrated contributions from all partners
1.0	28.06.2022	Final D6.3 version for submission
2.0	11.11.2022	Update (addressing privacy concerns in UC1)
2.5	30.1.2023	Update (addressing requirements from Ethics Check Report)
3.0	31.3.2023	Resubmission

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) 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. For testing the NLP components and pipelines of the SELMA platform, **SELMA Basic Testing and Configuration Interface** (Use Case 0) has been additionally introduced. It is used as both an internal testing platform and a public demonstration platform of the SELMA components and pipelines.

Table of Contents

- Executive Summary*..... 3**
- 1. Introduction*..... 6**
- 2. Types of Data Collected* 7**
 - 2.1 Data Types8**
 - 2.2 Requirements for Monitoring Data (UC1).....9**
 - 2.3 Requirements for News Production Use Case (UC2) 10**
 - 2.4 Requirements for the Open-Source Platform (UC0) 10**
 - 2.5 Requirements for Technology-Specific Data 10**
 - 2.5.1 Raw Data and Metadata 11
 - 2.5.2 Transcribed Data 12
 - 2.5.3 NER Annotated Data 14
 - 2.5.4 Entity Linking Data 15
 - 2.5.5 Topic Detection Data 16
 - 2.5.6 Storyline Clustering Data 16
 - 2.6 Provision of Monitoring Data 16**
 - 2.7 Provision of Technology-Specific Data 19**
 - 2.7.1 Transcribed Data 21
 - 2.7.2 Annotated Data 22
 - 2.8 User data 24**
- 3. Types of Generated Data* 27**
- 4. Data and Metadata Standards* 28**
 - 4.1 Data Identifiers and Internal Data Format 28**
 - 4.2 Text Feeds..... 29**
 - 4.3 Audio & Video Feeds..... 29**
 - 4.4 Entity Identifiers and Properties..... 30**

5. Data Storage, Preservation, Reuse and Sharing	31
6. Policies for Data Access and Sharing	34
7. Conclusion	36

Table of Figures

Figure 1 A sample script in the markdown format.....	22
Figure 2 Sample NER-annotated data	23
Figure 3 A JSON data snippet illustrating the SELMA internal data exchange format	29

Table of Tables

Table 1 Amount of audio/video data provided by DW for technology development.....	20
Table 2 Data fields provided for each DW news bulletin script	21

1. Introduction

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

- An updated description of what research and innovation activities of the project use which data, and a description of who is 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 are being put in place.
- Identification of 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 Data Management Plan also 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), actions taken by SELMA to reduce the identified risks.

This is the interim version of the Data Management Plan (cf. D6.1). This document is updated within the course of the project's development. A major update has been done addressing the report from the external ethics advisors in March 2023. There will be one more iteration (D6.5) which will make the final elaboration 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, speech translation, named entity recognition and linking, text classification, clustering, spoken language understanding, text and speech 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, 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 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 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 Tool (Use Case 2) – a multilingual editorial environment for journalists.

Both DW and Priberam target the 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, there has been introduced also Use Case 0 (UC0) – SELMA Basic Testing and Configuration Interface which is maintained by IMCS and is used internally by the SELMA partners for testing the language processing components and pipelines of the SELMA platform.

UC0 is both open-source¹ and publicly available² and, thus, is also used as a public SELMA demonstration platform. UC0 does not ingest data from external sources, only the data provided by its users/testers, and this data is not stored, and no logging is done with user actions.

“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 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,
 - audio & video.
- delivery type:
 - batch data (incl. text streams, no audio live streams);

¹ <https://github.com/SELMA-project/SELMA-project.github.io>

² <https://selma-project.github.io/>

- 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., neural language models,
- user data which is needed to ensure restricted access to the MONITIO and *plain X* platforms is securely processed and stored in order to comply with GDPR,
- public datasets used in the literature and datasets licensed by each of 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 different 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 Use Case (UC2)

Use Case 2, the news production platform, is a tool platform 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 to debug and disseminate the technologies developed within 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 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.

The members of the consortium are aware that the kind of AI models produced are conditioned by the bias contained in the training datasets. It's the responsibility of each researcher to address and assess the quality of the datasets used. The Consortium in the Ethics deliverable (D8.1) produces an assessment on the output models, based on the datasets used for training, the risks and their possible dual-use.

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 within the scope 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, as well as Russian, Ukrainian and Latvian.

2.5.1 Raw Data and Metadata

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

For training entity representations, we are using the Wikipedia and Wikidata dumps. Additionally, to take into account entity drift and being able to populate the knowledge base, an online mechanism is being studied to update these representations. For this, we use all data we have from the scrapping news web sites. This data is stored in Priberam internal servers. Researchers within 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).

To reuse the same datasets also 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 can be 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 (with regard to the length of the clip).

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 dataset³ 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 can be used to ingest articles and their teasers from the DW archive. As a fallback solution to training and evaluation data collection for abstractive summarization (in case of less-resourced languages), the lead paragraphs of news articles can be regarded as implicit summaries.

2.5.2 Transcribed Data

For the automatic speech recognition and transcription (ASR), in addition to the raw audio/video datasets (Section 2.3.1), already existing datasets of transcribed speech corpora will also be reused within the SELMA project. No additional datasets of transcribed speech will be created for the ASR development.

For the automatic text-to-speech synthesis (TTS), however, transcribed speech datasets of a limited amount will be created for some languages (e.g., Brazilian Portuguese and Latvian) 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;

³ https://huggingface.co/datasets/cnn_dailymail

- i. open-access datasets like the multilingual M-AILABS⁴, CSS10⁵ and CommonVoice⁶ speech datasets, the TED-LIUM3 speech dataset⁷, and the very recent Spotify Podcast Dataset⁸ will be considered;
 - ii. restricted-access datasets like QUAERO and ETAPE are available for internal use;
- 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 (at least 11.6k hours of diverse recordings; more details in table 1 in section 2.5) 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 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 transcribed speech corpus, a correct (curated) 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. For the speech summarization needs, a text teaser is required instead of full transcription.

⁴ <https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/>

⁵ <https://github.com/Kyubyong/CSS10>

⁶ <https://commonvoice.mozilla.org>

⁷ <https://www.openslr.org/51/>

⁸ <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 has to be clarified between the data provider and the technical partners.

2.5.3 NER Annotated Data

For named entity recognition (NER) creation of a multilingual dataset is in progress by the SELMA consortium. For a selected subset of SELMA languages for which compliant prior datasets are not available, a representative set of approximately 500–1000 documents (news articles) per language can be semi-automatically annotated by SELMA partners according to a common NER annotation schema. Since the manual annotation process is very time-consuming, the whole annotation process can be speeded up by manually annotating a few hundred documents, then training a neural model for automatic NER followed by manual curation of automatically pre-annotated remaining set of documents. The total amount of annotated/curated documents can be reduced by using the language transfer mechanisms researched within the project (promising experiments have already been conducted by joint training with high-resourced German and less-resourced Latvian data).

Regarding the document selection for each language, the focus is 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 random public persons (mentions of person names and related entities). Nevertheless, the set of selected articles for each language should be diverse (representative) in terms of topics, time periods, authors, channels. Therefore, datasets for NER annotation can be partially collected from DW news feeds, but they should be 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) is used as the fundament and orientation for the SELMA multilingual dataset.

The 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 therefore will be shared with an open license (e.g., CC BY 4.0), 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. For the rest of the languages, the copyright protection of the content (news articles) itself can be ensured by scrambling the datasets before making them public (e.g. by reordering the sentences in all the articles in an alphabetical order). This sharing of the dataset is still under consideration and waiting for legal advice on whether it can or not be made public given the description above.

Since SELMA partners use private GitHub repositories for development purposes, the sharing of the open NER dataset will be done via a public GitHub repository that then has to be disseminated to reach a wider research community. Additionally, the dataset will be distributed via the European Common Language Resources and Technology Infrastructure CLARIN⁹. This will ensure not only sustainability of this essential language resource, but will also facilitate its discovery, reuse and citation within the language technology community.

2.5.4 Entity Linking Data

For entity linking the project is using 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 being done on UC1 monitoring data.

⁹ <https://www.clarin.eu/content/services>

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 dataset¹⁰ 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 Finish dataset contains articles from 1992 to 2018. This new dataset together with the previously licensed LUSA dataset will enable us to further explore on the multilingual classification task and expand the diversity of the dataset.

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

2.5.6 Storyline Clustering Data

For the online clustering training and evaluation data, the project is using the dataset provided by Rupnik et al.¹¹ in “cross-lingual document similarity and event tracking” .

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

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 currently being collected from Twitter and YouTube channels for selected media providers through the ingestion pipeline. The platform is currently ingesting about 300 thousand news articles per day.

¹⁰ STT. Finnish News Agency Archive 1992-2018, source [text corpus]. Kielipankki. Retrieved from <http://urn.fi/urn:nbn:fi:lb-2019041501>

¹¹ Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., Grobelnik, M.: in News across languages - cross-lingual document similarity and event tracking. In: Journal of Artificial Intelligence Research, Special Track on Cross-language Algorithms and Applications.

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 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 the case of DW, full content and metadata ingestion through the DW proprietary API is also being done for a better data quality (in comparison to scraping). See Section 4.2 for more details.

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 is collected and ingested into the SELMA platform. News items from other public media sites are being 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 publishers are increasingly publishing unique content on social media platforms like Twitter, Facebook, Instagram, TikTok and YouTube, we have applied for access to gather data from **public media publishers pages** from Facebook, Instagram and Twitter with success. 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 benefit even for using aggregated user data from social media. As such, the adopted policy for collecting data from 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 Wikidata¹² knowledge graph hosted by Wikimedia Foundation. For each processed article (a news item), the automatic named entity linker (NEL) of the SELMA platform 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. The linked entities and their properties are stored and indexed in a database of Use Case 1. 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 some sensitive personal data categories for the diversity monitoring. Nevertheless, it might still rise some potential ethical issues. To mitigate such potential ethical risks, the SELMA project closely and regularly evaluates the development of the Use Case 1 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 Use Case 1 and in the commercial platform. The ethical concerns and mitigation of risks are addressed in more detail in D8.1 (Section 9 in particular).

The system will try to cover as much of the published media as possible.

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 the respect of copyright, media monitoring content data is only shown to users of the platform when Priberam has in place an agreement with the Publishers or their representative

¹² https://www.wikidata.org/wiki/Wikidata:Main_Page

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);
- free 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 will be accessible only to that entity (client or partner).

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

Restrictions on content use and visibility within the platform and the Consortium:

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

2.7 Provision of Technology-Specific Data

In order to develop specific technology components, the consortium both annotates new data and collects existing data from its internal repositories. DW provides data upon request when such data is available (e.g., raw audio & video data, transcribed speech data for TTS, news articles and their 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 platforms, additional data will be gathered via user feedback. User feedback data will be used at least for entity linking and retrieval modelling.

To mitigate the risk of potential bias, discrimination and stigmatisation in the output models we strive 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.

For training wav2vec speech recognition models for the selected SELMA languages (Section 2.3.1), lists of DW audio/video recordings are collected by DW and provided to the SELMA technology partners. This is 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 16 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) currently collected and provided for the development of language processing components of the SELMA platform (more than 11.6k hours in total).

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

Language	Amount (hours)	Language	Amount (hours)
Arabic	116	Hindi	255
Brazilian	254	Indonesian	207
Chinese	396	Pashto	155
Dari	122	Persian	465
English	433	Polish	126
French	257	Russian	1,184
Greek	110	Turkish	1,306
Hausa	5,756	Ukrainian	518

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, only the DW data is used.

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

2.7.1 Transcribed Data

One specific dataset has been identified and created so far for training a TTS model, 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 June 2022, 870 audio bulletins have been collected, which results in approximately 5220 minutes (87 hours) of audio data. It should be noted that the Brazilian Portuguese dataset contains data from eight speakers – DW Brazil news announcers. For the first iteration of the TTS model, generated at the end of 2021, the number of collected hours per speaker ranged from 0.7 to 8.5 hours.

In addition to the audio files, there is an automatically generated subtitle file (SRT) available for each bulletin. For most of the bulletins (767 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 2 and Figure 1).

Table 2 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 and ingested 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

Additional 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. In addition, 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.

2.7.2 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.

A mix of priority languages, both high-resourced and low-resourced, have been selected for DW annotation. Initially, this includes Arabic, Pashto, Serbian and Turkish. Other languages are considered to be added, including Chinese (both versions), Greek, Hindi, Indonesian, Persian, and Polish. 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 (about to start for Turkish) and IMCS for Latvian (about to start for Russian and Ukrainian). More languages will be gradually annotated and added to the dataset.

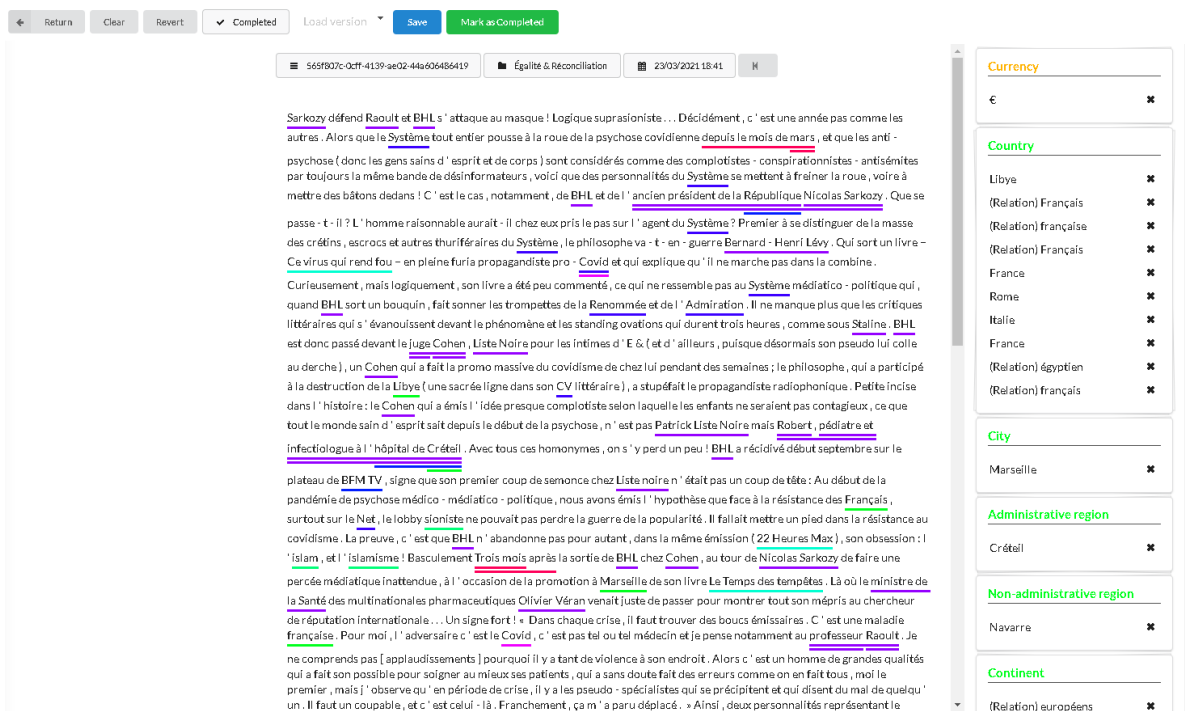


Figure 2 Sample NER-annotated data

As our experience with the highly inflected Latvian language shows, after manual NER-annotation of ~750 articles using the fine-grained Priberam Named Entities Annotation Schema, it was possible to train an automatic NER tagger with ~85% accuracy. This has allowed IMCS to switch from manual annotation to manual curation, which significantly improves productivity. The same approach will be applied to Russian, Ukrainian and all the DW selected languages.

Another observation is that it is not sufficient to collect articles for NER-annotation from a single source like DW. To make the datasets diverse and representative, we start with an initial set of at least 500 articles (per language) collected via the DW API (selecting medium size articles of various topics from various time frames) for the manual annotation phase and will

continue with articles collected from various external sources in the automatic pre-annotation and manual curation phase. To protect the copyrights, these datasets will be scrambled at paragraph or sentence level before releasing them as open data, as outlined in Section 2.5.3.

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, that 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 into 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.¹³
- “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

¹³ Accessible at [https://priberam.pt/Docs/Priberam Política de Privacidade.pdf](https://priberam.pt/Docs/Priberam%20Politica%20de%20Privacidade.pdf)

in the platform and the data generated in the course of 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 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 guidelines¹⁴ 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. The project aims to use this data to improve the results in certain tasks by automatic post

¹⁴ <https://pages.nist.gov/800-63-3/sp800-63b.html>

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;
- 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 during the project:

- Content data generated during media monitoring (Use Case 1) and news production (Use Case 2), as well as testing of the SELMA platform (Use Case 0). This is typical broadcast data that remains copyright-protected. See Figure 1 (Section 2.5.2) 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 2.5.3) 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 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 fahren 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 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 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 Zelenskyy¹⁵).

Following our decision (see Section 2.4) not to collect and analyze any highly 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), ~~country of citizenship (P27)~~, ~~occupation (P106)~~, ~~ethnic group (P172)~~, ~~sexual orientation (P91)~~, ~~religion or worldview (P140)~~, ~~medical condition (P1050)~~, ~~educated at (P69)~~, ~~member of political party (P102)~~, continent (P30; to support analysis by this property only at the Global South vs. Global North level).

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.

¹⁵ <https://www.wikidata.org/wiki/Q3874799>

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 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; see Section 2.5.3).

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 platform:

- github.com/SELMA-project/brasil-noticias-scripts – contains scripts of news bulletins produced by DW Brazil, together with links to the respective audio/video files that are publicly available from DW and YouTube websites, cannot be released as open data (see Section 2.5.2);
- github.com/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 Section 2.5.1);
- github.com/SELMA-project/youtube-audio-data – additional lists of audio/video data, publicly available from YouTube, cannot be released as open data (see Section 2.5.1);
- github.com/SELMA-project/HNNER_Torch – NER-annotated datasets created within the project, will be released as open data (see Section 2.5.3);
- datasets being created for training NER and NEL models are currently 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 and IMCS are planned to be released as open data (see Sections 2.3.3 and 2.5.3).

- ingested and annotated data on UC1 are stored in a private Postgres database at Priberam premises. These data will not be released outside of the project, 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 to 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 use selected datasets (like the Brazilian Portuguese dataset described in Section 2.5.2) 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 are retrieved from the DW repository and stored on the partner servers.

The technical partners retrieve the technology-specific data from the shared repository and 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 Priberam, and, consequently, ingested content will be 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 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. It should be noted that any cloud services that might be selected for the use within SELMA will always be located in EU to respect GDPR.

Data preservation and sharing options after the project will be discussed in the final version of the Data Management Plan, when the final output form of data will become visible. However, it is envisaged that Use Case 0 – the open-source SELMA orchestration and testing 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.
- 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.

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 any social-media-user (comments, replies etc..) except for the news publisher social-media-user with whom we have the license agreement. 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 to the completion of the project will be stored.

Data security procedures will be established for each partner dealing with SELMA datasets. Regarding the non-highly-sensitive personal data Wikidata properties of named entities such as binary gender, collected and stored to support the SELMA diversity use case (cf. D1.1), mitigation procedures of potential ethical issues are described in D8.1 (Section 5.8 in particular).

Access to the SELMA data repositories and to the SELMA platform (populated with data) is secured using SSL via the HTTPS protocol and will require authentication (except for the public SELMA demonstration 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 employees. 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.

Aside from the dispositions within this document, all consortium partners dealing with data, including provision, use, processing and storing, 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 interim Data Management Plan (a significantly updated and extended version of 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 and aspects 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 are gradually made available as open and accessible to the research community, when and if such release is not in breach of copyright and personal data protection.

This is the second of three iterations of DMP. Major changes in the interim DMP (Sections 2.4, 4.4, and 6) reflect the major changes in Use Case 1 to resolve the identified ethical and personal data issues:

- We have arrived at the conclusion that the research being done in the scope of the project would not benefit for using aggregated user data from social media, therefore not even aggregated user data will be collected from social media;
- We have abandoned our initial plan to collect certain “sensitive” personal data Wikidata categories for the in-depth diversity monitoring; only few not highly-sensitive personal data categories (including binary gender) will be collected to support a very general diversity monitoring. This data will only be made available in a clustered view;

We have also added a detailed section (2.8) on the collection, storage and protection of personal data of the users of Use Case 1 and Use Case 2 platforms.

The final DMP version is due at the end of the project (M36). Data collection, generation and processing are key areas in the SELMA project and are discussed, elaborated and further specified throughout the project.