Cost effective, Multilingual, Privacy-driven voice-enabled Services

www.compriseh2020.eu

Call: H2020-ICT-2018-2020
Topic: ICT-29-2018
Type of action: RIA
Grant agreement Nº: 825081

WP Nº1: Project management

Deliverable Nº1.6: Final Data Management Plan

Lead partner: INRIA

Version Nº: 1.0

Date: 30/11/2021

Document information	
Deliverable Nº and title:	D1.6 – Final Data Management Plan
Version Nº:	1.0
Lead beneficiary:	INRIA
Author(s):	Akira Campbell (INRIA), Emmanuel Vincent (INRIA)
Reviewers:	Raivis Skadiņš (TILDE), Dietrich Klakow (USAAR)
Submission date:	30/11/2021
Due date:	30/11/2021
Type[1]:	ORDP
Dissemination level[2]:	PU

Document history			
Date	Version	Author(s)	Comments
05/10/2021	0.1	Akira Campbell	Draft deliverable
08/11/2021	0.2	Akira Campbell	Initial version
30/11/2021	1.0	Akira Campbell & Emmanuel Vincent	Final version integrating feedback and comments from the reviewers

[1] **R**: Report, **DEC:** Websites, patent filling, videos; **DEM:** Demonstrator, pilot, prototype; **ORDP:** Open Research Data Pilot; **ETHICS:** Ethics requirement. **OTHER:** Software Tools

[2] **PU:** Public**; CO:** Confidential, only for members of the consortium (including the Commission Services)

Document summary

“Project management” is the first Work Package (WP1) of COMPRISE. It aimed to plan, organize, and control all activities and tasks so that the project could run successfully throughout all stages. The activities involved in WP1 included setting up efficient collaboration tools to foster communication between the project partners, controlling risks, managing data, and monitoring quality, managing budget and related issues, and reporting to the European Commission. These basic elements contributed to achieving the project goals in a timely manner through efficient coordination. As the project coordinator, INRIA is carrying out most of the activities performed in WP1, while all other partners supported this effort.

So far, five WP1 deliverables were submitted at M2, M3, M6 and M12. The first deliverable describes the “Project quality plan and private web platform” (D1.1 − M2), the second deliverable explains the “Detailed work plan” (D1.2 − M3), the third deliverable explains the “Initial data management plan” (D1.3 – M6), the fourth deliverable explains the “Interim progress report” (D1.4 – M12) and the fifth deliverable explains the “Advisory Board assessment” (D1.5 – M12). In the current document, the last deliverable of WP1 the “Final data management plan” (D1.6) is presented.

D1.6 explains how the COMPRISE consortium has planned to manage the data gathered during the project. This includes how data was collected, used, managed, stored, sustainably archived, and disseminated.

Table of contents

1 Introduction

The Data Management Plan (DMP) describes the data management lifecycle for the data that has been collected, processed, used, managed, stored, sustainably archived, and disseminated over the course of COMPRISE. As part of the project's activities, the DMP has been a key element for efficient data management and it has helped the consortium partners make their research data Findable, Accessible, Interoperable and Reusable (FAIR).

This document presents the final version of the DMP submitted to the European Commission on M36 of COMPRISE.

D1.6 is prepared with respect to the European Commission guidelines and template dedicated to H2020 projects participating in the extended Open Research Data Pilot (ORD Pilot):

- H2020 Annotated Model Grant Agreement – Open access to research data[3]
- Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020[4]
- Guidelines on FAIR Data Management in Horizon 2020[5]
- FAIR Data Principles[6]
- The FAIR Guiding Principles for Scientific Data Management and Stewardship[7]
- Template Horizon 2020 Data Management Plan (DMP)[8]
- OpenAIRE Research Data Management Briefing Paper – Understanding Research Data Management[9]
- Digital Curation Centre (DCC) – Checklist for a Data Management Plan[10].

The rest of this document is structured as follows. Section 2 presents a summary of the data that has been managed over the course of COMPRISE. Section 3 details FAIR data and explains how to make data findable, openly accessible, interoperable, and how to increase data reuse. The details tied to the allocation of resources are provided in Section 4. Data security and ethical aspects are discussed in Sections 5 and 6, respectively. Finally, a conclusion is given in Section 7.

2 Data summary

This section provides a description of the datasets collected and used in COMPRISE. This will include the purpose and utility of the defined datasets as well as a technical description of these.

---

[3] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf#page=243 (Version 5.1, December 2018)
[4] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (Version 3.2, March 2017)
[5] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (Version 3.0, July 2016)
[6] https://www.force11.org/fairprinciples (Version B1.0)
[7] https://www.nature.com/articles/sdata201618.pdf (Article in nature (2016): DOI: 10.1038/sdata.2016.18)
[8] http://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx (Version 3.0, July 2016)
[9] https://www.openaire.eu/briefpaper-rdm-infonoads/view-document (April 2017)
[10] http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf (Version 4.0, 2014)

2.1 Data purpose and data utility

As previously defined in Deliverable D1.2 “Detailed work plan” (submitted to the European Commission on February 28, 2019 − Confidential), common research datasets have been defined and/or collected within WP6 “Evaluation and demonstration for practical use cases” for the new tools developed in WP2 “Privacy-driven voice interaction”, WP3 “Multilingual personalised voice interaction”, and WP4 “Cost-effective multilingual voice interaction”. Additionally, data has been gathered for the development and evaluation of the demonstrators, i.e., e-commerce, consumer, and e-health applications corresponding to real use cases.

The definition of common datasets has allowed the scientific validation of the project’s components and tools in combination with each other, and their evaluation within the demonstrators. More specifically, to train and evaluate state-of-the-art machine learning based models for speech-to-text and dialogue processing, large amounts of training data are required. To this end, publicly available corpora have been assessed and acquired for the specific following purposes:

- speech-to-text,
- speech and text anonymisation,
- spoken language understanding and dialogue management,
- machine translation and its interaction with speech-to-text and text-to-speech,
- robust integration of machine translation and dialogue.

Whenever needed, these corpora have been complemented by the collection of small domain-specific corpora for research purposes. This resulted in a set of training and testing datasets that have been used in all tasks throughout the project. In addition, the components and tools developed in COMPRISE have also been evaluated on the data gathered as part of the demonstrator development and evaluation process. This has provided feedback to the demonstrator developers. Further details tied to the objectives of each dataset used during COMPRISE are given in Appendix B “Dataset forms”.

2.2 COMPRISE datasets

As mentioned in Deliverable D1.3 “Initial data management plan” (M6) the consortium members identified several datasets to be used and collected for each specified aim. Some of the datasets defined for research purposes have been used in their current state (see Appendix A “Unrevised datasets”), while some others have been enriched with additional collected data. As for the demonstrators, some existing datasets have been enriched with extra data, while others have been specifically built from scratch for each demonstrator’s use case.

In the following, we only discuss some of the existing datasets that have been modified and/or enriched within COMPRISE and datasets that have been collected for evaluation. A summary of these is given in **Table 1**, whereas a full detailed description is provided in Appendix B “Dataset forms”. A detailed description of the unmodified datasets falls out of the scope of this document.

Table 1: List of COMPRISE datasets.

Datasets for research purposes					
Identifier	**Title**	**Partner**	**Data type**	**Access**	**WP**
COMPRISE_Data 01_Drive-Thru_V1.0	Drive-Thru beta testers dataset	NETF	Text and relational data	Confidential	WP6
COMPRISE_Data 02_TBTS_V1.0	Tilde Balss Test Set	TILDE	Speech & Text & Translation	Confidential	WP3
COMPRISE_Data 03_LibriSpeech_ V1.0	Librispeech[11]	INRIA	Speech & Text & Speaker ID annotations	Public	WP2 WP4
COMPRISE_Data 04_CommonVoic e_V1.0	Mozilla Common Voice[12]	INRIA	Speech & Text & Speaker ID annotations	Public	WP2
COMPRISE_Data 05_VerbMobil-1Corpus_V1.0	VerbMobil-1 Corpus[13]	USAAR	Audio & Transcripts & Annotation	Public but paid	WP2 WP3 WP4
COMPRISE_Data 06_Yelp_V1.0	Yelp Dataset[14]	USAAR	Text & Annotation	Public	WP2
COMPRISE_Data 07_HYTS_V1.0	Hausa and Yoruba Text classification[15]	USAAR	Text & Annotation	Public	WP4
COMPRISE_Data 08_H-VOA-NER_V1.0	Hausa VOA NER[16]	USAAR	Text & Annotation	Public	WP4
COMPRISE_Data 09_H-VOATopics_V1.0	Hausa VOA Topics[17]	USAAR	Text & Annotation	Public	WP4
COMPRISE_Data 10_MENYO-20K_V1.0	MENYO-20L[18]	USAAR	Text & Translation	Public	WP3
COMPRISE_Data 11_M-NER_V1.0	MasakhaNER[19]	USAAR	Text & Annotation	Public	WP4
COMRPISE_Data 12_Y-GV-NER_V1.0	Yoruba GV NER[20]	USAAR	Text & Annotation	Public	WP4

[11] DOI: 10.5281/zenodo.5736583
[12] DOI: 10.5281/zenodo.5736808
[13] https://github.com/uds-lsv/privacy-preserving-text-transformer
[14] https://github.com/uds-lsv/author-profiling-prevention-BT
[15] https://github.com/uds-lsv/transfer-distant-transformer-african
[16] https://github.com/uds-lsv/transfer-distant-transformer-african/tree/master/data/hausa_ner
[17] https://github.com/uds-lsv/transfer-distant-transformer-african/tree/master/data/hausa_newsclass
[18] https://github.com/uds-lsv/menyo-20k_MT
[19] https://github.com/masakhane-io/masakhane-ner
[20] https://github.com/ajesujoba/YorubaTwi-Embedding

Datasets for research purposes					
Identifier	**Title**	**Partner**	**Data type**	**Access**	**WP**
COMPRISE_Data 12_Y-BBCTopics_V1.0	Yoruba BBC News Topics Classification Dataset[21]	USAAR	Text & Annotation	Public	WP4
COMPRISE_Data 13_Notes-ED_V1.0	Notes Evaluation Dataset	ASCO	Speech, text & Annotation	Confidential	WP6
COMPRISE_Data 14_Shoplay-ED_V1.0	Shoplay Evaluation Dataset	NETF	Speech & Text	Confidential	WP6
COMPRISE_Data 15_HC-ED_V1.0	Hospital Concierge Dataset	TILDE	Speech, text & Annotation	Confidential	WP6
COMPRISE_Data 16_DA-ED_V1.0	Doctor's Assistant Dataset	TILDE	Speech, text & Annotation	Confidential	WP6

During the process of COMPRISE, sixteen datasets have been modified, enriched, or collected to conduct the research of the project. Four of these datasets are the direct result of the efforts of the project. These demonstrator evaluation datasets have been collected with the demonstrators during their development with the aim to understand the capabilities and weaknesses of each app. These datasets contain a rich source of real user utterances which not only allowed the evaluation of the components, but also the possibility to understand how users actually interact with voice-enabled systems, such as the various intuitive actions that may differ from what the users initially thought they wanted.

Aside from the collection of data from the demonstrator apps, other datasets were collected to evaluate the methodology of working on small and limited data languages. These datasets demonstrated that automated data privacy is not limited to big data languages, and that they can be applied to a wider range of languages. These datasets are not only valuable to the evaluation of methodology, but also provide long-term utility for the scientific and industrial community within Europe and afar.

The other datasets identified for research purposes are considered as publicly available benchmark datasets. Though these datasets consist of general conversations and do not contain explicitly privacy-related annotations, such ground-truth data was required, and additional annotation has been added for the evaluation of the privacy-driven speech and text transformations in COMPRISE. Additionally, for approaches based on supervised machine learning, data annotations are required for training the classifiers. Existing metadata on the speaker's gender, age, and possibly profession was available and has been used to this end. Such annotation data that enriches the value of these datasets have been made public, as have the training, development and testing data categorisation that have been used. By making these items public, the scientific community will be capable of replicating, evaluating and improving on the results

[21] https://github.com/uds-lsv/transfer-distant-transformer-african/tree/master/data/yoruba_newsclass

achieved by COMPRISE and further advance the domain of privacy conscious but cost-effective language development. The utility for each defined dataset is detailed in Appendix B “Dataset forms”.

2.3 Data technical details: origin, type, format, and size

The data that was used during the course of COMPRISE, and that was gathered by different partners, has been both quantitative and qualitative in nature. It was analysed from a range of methodological perspectives for project development and scientific purposes.

As previously mentioned, many of the used datasets have been made openly accessible, while some of the collected datasets remain confidential. For all openly accessible datasets, and to maximize their interoperability, management, and reuse, the consortium members used, whenever possible, formats that are non-proprietary, unencrypted, uncompressed and in common usage by the research community. To this end, the consortium members followed whenever possible the indications of the UK Data Archive,[22] as recommended by OpenAIRE,[23] and as displayed in **Table 2**.

Table 2: File formats recommended by OpenAIRE.

Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata Variable labels, code labels, and defined missing values	SPSS portable format (.por) Delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) Structured text or mark-up file of metadata information, e.g., DDI XML file	Proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.acc)
Tabular data with minimal metadata Column headings, variable names	Comma-separated values (.csv) Tab-delimited file (.tab) Delimited text with SQL data definition statements	Delimited text (.txt) with characters that are not present in the data used as delimiters Widely used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)
Textual data	Rich Text Format (.rtf) Plain text, ASCII (.txt) EXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html) Widely used formats: MS Word (.doc/.docx) Some software-specific formats: NUD*IST, NVivo and ATLAS.ti
Audio data	Free Lossless Audio Codec (FLAC) (.flac)	MPEG-1 Audio Layer 3 (.mp3) if original created in this format Audio Interchange File Format (.aif) Waveform Audio Format (.wav)

[22] https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats
[23] https://www.openaire.eu/

Type of data	Recommended formats	Acceptable formats
Documentation and scripts	Rich Text Format (.rtf) PDF/UA, PDF/A or PDF (.pdf) XHTML or HTML (.xhtml, .htm) OpenDocument Text (.odt)	Plain text (.txt) Widely used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) XML marked-up text (.xml) according to an appropriate DTD or schema, e.g., XHMTL 1.0

Besides these recommended formats, the consortium agreed on an additional data format necessary for the demonstrator datasets: JavaScript Object Notation (JSON).

During the course of COMPRISE, a variety of results were obtained from the various work packages. The data related to these results have been made public by means related to the ease of access and the vast outreach of the platforms Zenodo[24] and GitHub. Aside from Zenodo, data have also been placed in GitHub due to the community connections that were created during the collection and creation of the data. For this reason, some of the data has been disseminated using Git repositories as indicated in the footnotes of **Table 1**.[25]

On the other hand, the confidential demonstrator datasets that have been collected during the course of the project have been stored in each responsible partner's storage facilities (see Section 5 "Data security").

3 FAIR data

In this section, a description of how to make the COMPRISE research data FAIR (findable, accessible, interoperable, and reusable) is provided.

3.1 Making data findable, including provisions for metadata

To make data findable, effectively, and persistently citable when it is uploaded to the Zenodo repository, a Digital Object Identifier (DOI) has been assigned to each uploaded open dataset as listed in **Table 3**. This DOI, which is automatically provided by Zenodo to all publicly available uploads, can be used in any relevant publications to direct readers to the underlying dataset. As for data that is in continuous active use, GitHub and GitLab repositories will remain the primary reference to keep the community interaction ongoing.

Table 3: Example of Digital Object Identifier (DOI) uploads

Type of data	DOI number
LibriSpeech training data splits	10.5281/zenodo.5736583
Mozilla training data splits	10.5281/zenodo.5736808
VerbMobil annotations	10.5281/zenodo.5742055

To further emphasise how to make data findable, this Deliverable D1.6 has relied on the collection of data management forms that have been filled out during the COMPRISE

[24] https://zenodo.org/communities/comprise/
[25] Datasets that are publicly available on GitHub will be linked to the COMPRISE Zenodo community for archiving in due course.

implementation progress. These forms have been filled by the project partners who are responsible for collecting data. These data management forms are only accessible via the project's internal communication tool named Partage while the data itself has been securely stored in either MyBox if it is to become publicly available, or in the responsible partner's storage facilities if it is to remain confidential.

The COMPRISE DMP versions have supported research reproducibility and trustworthiness, as they have emphasised the need to accurately cite and identify the exact version of the dataset used as research input and underpins the research findings. The project's DMP is based on a consistent version numbering scheme[26] that has enabled the consortium members to track changes within the data collection, to determine precisely which version was used previously and which version is currently under use, and to set expectations about how each version differs from the previous one.

All the used data will continue to be identified by a versioning indicator and a history. The opted versioning scheme is based on a two-part numbering format:[27] Major.Minor (e.g., V2.3). The "Major" data revision part indicates a significant change in the content of the dataset that may bring substantial modifications in the scope, in the context or in the intended use of the dataset. For instance, a "Major" data revision reflects the following situations:

- addition/deletion of new data items to/from a collection,
- introduction of an additional set of data features,
- modification of the format of the data items.

As for the "Minor" data revision part, it involves quality improvement over existing data items. "Minor" changes may not affect the scope, or the context or even the intended use of the initial data collection. This part has reflected the following situations:

- renaming of data features,
- correction of errors in the existing data collection,
- rerunning of a data generation model with adjustment of some parameters.

This versioning scheme has guaranteed specificity and verifiability and has enabled each version of a given data collection to be uniquely referenced.

These versions have been saved as a compressed file, e.g., a ZIP file. An editable copy of the latest versions has also been saved to allow easy revision. As previously mentioned, only the collected public datasets have been stored in MyBox, and hence their revisions in both file formats have also been stored in this repository. Whenever a change or an update was made to the documents in MyBox, the concerned COMPRISE members have been notified. The "confidential" and "public but paid" datasets have been stored in both formats on the partner's own storage facilities.

The above schema has been used in COMPRISE to establish a naming convention for the project datasets. This consists of the following items:

- a prefix "COMPRISE" indicating that the dataset was prepared in the course of COMPRISE,

[26] W3C Data on the Web Best Practices guide. https://www.w3.org/TR/dwbp/#dataVersioning
[27] The Australian National Data Service. https://www.ands.org.au/working-with-data/data-management/data-versioning

- "Data", referring to dataset, followed by a unique chronological number of the project overall collected datasets,
- the title of the dataset,
- a versioning number based on the above versioning scheme.

For instance, the first project dataset has been identified as: "COMPRISE_Data1_VerbMobil-1Corpus_V1.0".

To further make data findable and to optimise the possibility for reuse, search keywords have been provided when the dataset was uploaded to Zenodo. As Zenodo follows the minimum Data Cite metadata standards, additional technical keywords have been added to each dataset.

Every dataset has been annotated with metadata that includes the identifier of the dataset. As COMPRISE relies on the Zenodo repository, it has benefited from the Zenodo JSON metadata schema[28] and Data Cite metadata standards that offer key data documentation such as:

- creators and their affiliations,
- data location and persistent identifier scheme,
- chosen license,
- funding,
- contributors,
- references,
- related journals, conferences, books and/or thesis,
- deposit metadata.

Adding to the Zenodo metadata schema, the uploaded datasets have been accompanied by a metadata file (see Appendix C “Metadata file template”) containing the information provided in the data management form given in Appendix B “Dataset forms”. Upon the update and/or a new upload of a public dataset, the information given in the data management form as well as in the metadata file has been updated and uploaded to the Zenodo repository.

3.2 Making data openly accessible

COMPRISE has enriched and collected a variety of datasets, which have different natures and access privileges. These different access privileges are described in detail in Appendix B “Dataset forms” and are reviewed here in a concise manner. Appendix B “Dataset forms” also provides a detailed description of all aspects related to dataset management. A short overview on data access policies and availability is presented in this section.

According to the long-term datasets utility (see Section 2.1 “Data purpose and data utility”) and potential limitations due to the protection of personal data (see Section 6 “Ethical aspects”), different levels of confidentiality have been considered within the project consortium:

[28] https://zenodo.org/schemas/records/record-v1.0.0.json

- **Confidential to partner.** This option is applied when the case is tied to a dataset that is collected by a specific partner and that contains personal data that cannot be protected once disclosed.
- **Confidential to consortium members (including the Commission services).** This option is applied for data containing confidential information or those with no wide scope of use and long-term value.
- **Public.** This option has been applied to most COMPRISE datasets.

Table 1 in Section 2.1 “Data purpose and data utility” highlights the publicly available research datasets that have been used in COMPRISE that have witnessed revisions by adding newly collected data. Two such types of datasets can be distinguished. The first type refers to datasets available at no cost while the second type refers to datasets available for a fee.

The first type of datasets is already openly available; thus, the newly collected data has also fulfilled their usefulness for scientific or other public purposes. Regarding the paid datasets, these have been kept in the storage space(s) of the consortium member(s) who purchased them, and hence considered as confidential to partner to protect personal information as described in “D8.1 − POPD − H − Requirement No. 1” (Submitted to the European Commission on May 31, 2019 − Confidential). These datasets have fulfilled the policies on data backup and preservation of the corresponding consortium member and will be maintained by this entity after the end of the project. However, the newly collected annotations for these datasets have been made publicly available, after negotiation with the copyright holder whenever required.

The datasets gathered as part of the demonstrator evaluation process have been classified as confidential to consortium members and be used for the restricted purpose of this project only. It is important to note that this confidentiality constraint has not impacted the dissemination of the project outcomes in terms of voice interaction technology. The consortium member in charge have archived the data as set out in Section 5 “Data security”. Like all confidential data in COMPRISE, its preservation and maintenance during and after the project has and will continue to be handled by the data owners.

As previously mentioned, to facilitate deposit, update, and management, the collected public data has been made available via the Zenodo COMPRISE created community.[29] As for publications, these have been made open access via the public repository HAL.[30]

Visibility and access to publicly shared datasets has been facilitated by Zenodo metadata and search facility as well as to the automatic link to both OpenAIRE and to the CORDIS project page.[31]

Moreover, to increase dataset accessibility and reusability, the consortium agrees to provide full software and tools information for all datasets within the documentation provided in the data management forms (see Appendix B “Dataset forms”).

Software plays a key role in COMPRISE and particular provisions should hence be considered for software development as part of the project activities in addition to provisions for access and rights agreed by partners in the Consortium Agreement. To preserve and share software code and documentation, the consortium members have

[29] https://zenodo.org/communities/comprise/
[30] https://hal.inria.fr/
[31] https://cordis.europa.eu/project/rcn/218720/factsheet/en

used Software Heritage[32]. Only open-source software has been publicly shared within this platform. However, software which is not publicly released has been either uploaded to GitLab,[33] if it is to be shared between the consortium members only, or in the partner's own repository if it is to remain private.

3.3 Making data interoperable

COMPRISE has collected and documented the data in a standardised way to ensure that the datasets can be understood, interpreted, reused, and shared in isolation alongside the accompanying metadata and documentation.

As previously described, all data collected in COMPRISE has been fully documented via the data management forms (see Appendix B "Dataset forms") and accompanied with detailed metadata supported by a set of select keywords (see Appendix C "Metadata file template"). This was to facilitate an automatic integration of COMPRISE data for other purposes allowing interdisciplinary interoperability. All data has been provided in generally used extensions, as described in Section 2.3 "Data technical details: origin, type, format, and size", adopting well established formats whenever possible which has also facilitated its reuse by other parties.

Standard vocabulary has been used for all data types present in the dataset to allow interdisciplinary interoperability. In addition, whenever required, the documentation has included a general glossary used to share information about the vocabulary and general methodologies employed for the generation of the dataset.

3.4 Increase data reuse

The collected public data has or will be made openly available. To allow the widest possible reuse, the consortium has attached a specific license to every deposited dataset. This has allowed the definition of all the work conditions as being under an open or a restricted access.

Zenodo automatically offers five different licensing options among Creative Commons (CC) Licenses, all foreseeing the attribution requirement to appropriately credit the authors for the original creation. Whenever possible, the Creative Commons Attribution 4.0 International (CC BY 4.0)[34] license has been used, to allow third parties to share and adapt data with no restrictions if attribution is provided.

In cases where the partner would like to further limit access to the uploaded data, an alternative license has been selected among the following options offered by Zenodo:[35]

- **Creative Commons Attribution Share-Alike 4.0 International (CC BY-SA 4.0).** Allows modification of the data for any purpose as long as it is distributed under the same original license (or a license listed as compatible).
- **Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0).** Allows distribution of the data for any purpose, but forbidding the distribution of derivative work.

[32] https://www.softwareheritage.org/
[33] https://about.gitlab.com/
[34] https://creativecommons.org/licenses/by/4.0/
[35] https://zenodo.org/record/1488616#.XLoeYOgzY2w

- **Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).** Allows sharing and modification, but limiting use to non-commercial purposes.
- **Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NCND 4.0).** Allows sharing but restricting both derivative work and commercial use of data.

Although not directly provided through Zenodo, an additional Creative Commons Attribution license can be applied upon specific request to the Zenodo team:

- **Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).** Allows modification as long as it is distributed for non-commercial purposes and under the same original license (or a license listed as compatible).

All public collected data has or will be stored in Zenodo, where it will remain available for the lifetime of the repository, which is currently warrantied for a minimum of 20 years. This will promote its reuse by other researchers and end-users, thereby contributing to the dissemination of the COMPRISE technological components and research advances.

4 Allocation of resources

As previously mentioned, a Zenodo repository was created at no cost for the project's open collected data, therefore ensuring data availability, backup, and versioning. Long-term preservation will be guaranteed for the lifetime of the Zenodo repository. This is currently the lifetime of Zenodo's operator, the European Organisation for Nuclear Research (CERN), which currently has an experimental programme defined at least for the next 20 years. After the end of that period, the collected data will be archived at the data owner's facilities.

As for the collected data tied to the confidential datasets, it will be managed by the partners responsible for its collection. Therefore, its maintenance, backup and versioning and long-term preservation and archival will be guaranteed by the partners' own resources and at their own expense.

The project coordinator oversaw the DMP from both the scientific and technical perspective. The registration of datasets and metadata, as well as backing up data for sharing through open access repositories, was the responsibility of the partner that gathers the data in its related work package. Quality control of these data was the responsibility of the relevant work package leader, supported by the Project Coordinator. Each partner has respected the policies set out in this DMP.

Publications featuring the data have been produced during the project and have been made available as open access on HAL, by selecting journals or conferences allowing immediate public access on institutional repositories, open access journals, or journals or conferences featuring a short embargo period.

5 Data security

During the course of COMPRISE, the collected data related to the public datasets has been stored in MyBox while the collected data tied to the confidential datasets has been stored in the responsible partner's storage facilities, as detailed in **Table 4**.

Table 4: Data storage description.

Partner institution/organisation	Data storage
INRIA	MyBox:[36] Seafile Professional Edition repository for secure sharing of large research data (e.g., audio files, model parameters, etc.). All project partners are members of the COMPRISE group, which can be accessed at https://mybox.inria.fr/#group/1059/. Each member can create one or more libraries (up to 10 GB) and view the data libraries created by others. Access to a given library can be restricted to any list of members and password protected. This makes it possible to share the data with different subgroups of partners, as appropriate.
USAAR	Data has been stored on local hard disks that can only be accessed through valid user accounts administered by USAAR. Non-local access through SSH (encrypted network connection) can be granted on request if necessary.
TILDE	Data has been stored on TILDE's servers which can only be accessed through valid user accounts administered by TILDE. Non-local access can be granted on request if necessary.
ASCO	Data has been stored on servers located within the European Union that can only be accessed through valid user accounts which are administered by ASCO. Non-local access can be granted on request if necessary.
NETF	Data has been stored on NETF's servers which can only be accessed through valid user accounts administered by NETF. Non-local access can be granted on request if necessary.

Regarding the collected data tied to the confidential datasets, different approaches have been used by each data owner's organisation, but common rules have applied. As presented in **Table 4**, datasets have been saved in servers and under the direct control and management of the organisation's personnel. Such infrastructure is equipped with different features, e.g., secure physical access, air conditioning, fire protection measures, hardware/electricity recovery measures, etc.

Different data access permissions, e.g., read-only, read-write, etc., have been granted to users and authorised computers by relevant staff, according to well-defined protocols. Additionally, confidentiality is guaranteed by supplementary methods, e.g., encryption and anonymisation, depending on the data's nature and applications. Furthermore, regular backups are envisaged for either security purposes, hardware failure recovery, or for archival purposes.

Following the completion of the project, all the responsibilities concerning data recovery and secure storage will go to the repository storing the dataset. Long-term preservation

[36] https://mybox.inria.fr/

is guaranteed even in the unlikely event that Zenodo will cease operation; migration of content on other repositories is planned.

6 Ethical aspects

The project's partners have complied with the recommendations set out in the Ethics Summary Report as well as with the ethical principles and standards under Horizon 2020 and relevant national, particularly, with the Regulation (EU) 2016/679 − General Data Protection Regulation (GDPR) − and international legislations, and any additional applicable laws of the member states concerned.

Indeed, with respect to the highest standards of research integrity, all partners have complied with the ethical principles as set out in the European Code of Conduct for Research Integrity;[37] these include, in particular, avoiding fabrication, falsification, plagiarism, or other research misconduct.

The COMPRISE project has focused on issues related to the collection and processing of data gathered through deep learning technologies and more specifically speech-to-text, spoken language understanding, and dialog management. Nevertheless, and with the aim of achieving a more satisfying user experience, massive amounts of data have been used, not only during the operating phase, but also during the training phase.

In this regard, the COMPRISE project have addressed the drawbacks that voice-controlled technologies entail in terms of costs, inclusiveness and, more specifically, privacy and ethical issues through the definition of a fully private-by-design methodology which, from a general point of view, means deleting as much personal information as possible from the users' speech thanks to privacy-driven transformations.

The other important objective of the COMPRISE project has been the creation of demonstrators which have covered different sectors and their evaluation by end-users to prove the benefits of voice-controlled technologies from the COMPRISE standpoint. End-users might provide some personal information and consequently, they have been properly informed of the processing tasks and the purpose of such processing so they can either accept it or reject it with full guarantees.

Finally, the corresponding explicit and written consent has been obtained to comply with the current legislation in the field. In the specific case of the e-health demonstrator developed by TILDE, some of these end-users have been patients in partnering hospitals. Additional measures have been taken in this case, in collaboration with doctors at the hospital such as the provision of an incidental findings policy.

Finally, it is obvious that human interaction and participation in this project has been essential to carry out the objectives of COMPRISE. However, it is to be noted that no physical interventions on such participants have been carried out. When statistical analysis, interviews or/and questionnaires have been conducted, personal information has been anonymised and personal data has been kept confidential.

All COMPRISE activities raising ethical issues have complied with the ethical requirements defined in Deliverable "D8.1 − POPD − H − Requirement No. 1" (Submitted to the European Commission on May 31, 2019 − Confidential). It includes informed consent forms, incidental findings, security measures, etc. The project partners have

[37] https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf (revised edition, 2017)

also complied with additional measures set out in Deliverable D5.1 “Data protection and GDPR requirements” (Submitted to the European Commission on May 31, 2019 − Public), and its revised version D5.5 “Final platform demonstrator and updated data protection and GDPR requirements” (Submitted to the European Commission on May 31, 2021 − Public).

7 Conclusion

Deliverable D1.6 presents the final version of the data management plan. Throughout this document, a list of datasets that have been collected, used, or modified is presented. These are accompanied by their technical description. Our implementation of how to make data findable, openly accessible, interoperable, and reusable through a clarification of licenses was also provided. Based on how to make data FAIR, the allocation of resources was described. Finally, a description of how to address data recovery as well as secure storage, together with a description of some other identified ethical aspects were also given.

Appendices

Appendix A Unrevised datasets

Some of the datasets defined for research purposes have been used in their current state, and hence have not witnessed any revisions, e.g., no additional data was collected to complement the datasets. **Table 5** presents a version of the defined used datasets that have not been revised. Most of these datasets are public.

Table 5: List of the unrevised datasets.

Identifier	Title	Partner	Data type	Access	WP
Saeima test set	Saeima test set[38]	TILDE	Speech & Text	Public	WP3
WMT 2017	WMT 2017[39]	TILDE	Parallel text	Public	WP3
Tildes Biura corpus	Tildes Biura corpus	TILDE	Text & Annotation	Confidential	WP3
IARPA Babel LT	IARPA Babel Lithuanian Language Pack[40]	TILDE	Speech & Text	Public but paid	WP3, WP6
LVDistated	Latvian Dictated Speech corpus	TILDE	Speech & Text	Paid	WP3, WP6
LVASR	Latvian Speech Recognition Corpus[41]	TILDE	Speech & Text	Paid	WP3, WP6
LIEPA	LIEPA	TILDE	Speech & Text	Public	WP3, WP6
DCEP	DCEP	TILDE	Parallel texts	Public	WP3
DG-TM	DG-TM	TILDE	Parallel texts	Public	WP3
Tilde Model	Tilde Model[42]	TILDE	Parallel texts	Public	WP3
Europarl corpus	Europarl corpus	TILDE	Parallel texts	Public	WP3
JRC-Acquis	JRC-Acquis	TILDE	Parallel texts	Public	WP3
Paracrawl	Paracrawl	TILDE	Parallel texts	Public	WP3

[38] http://metashare.tilde.com/repository/browse/asr-evaluation-corpus-from-data-of-saeima-of-the-republic-of-latvia/f222c75cfbdf11e9aa3b001dd8b71c66baa9439de9a945c8aa3bc741b749f29e/
[39] http://statmt.org/wmt17/
[40] https://catalog.ldc.upenn.edu/LDC2019S03
[41] http://www.lrec-conf.org/proceedings/lrec2014/pdf/284_Paper.pdf
[42] http://metashare.tilde.com/repository/browse/tilde-model-multilingual-open-data-for-eu-languages/a70af9701f6811e7aa3b001dd8b71c6645b57ed010b84385868c128cdff5807a/

Identifier	Title	Partner	Data type	Access	WP
Let's Go	Integral Let's Go[43]	INRIA	Speech & transcription	Public	WP4
VoxCeleb	VoxCeleb 1 & 2[44]	INRIA	Speech & Text & Speaker ID annotations	Public	WP2
AMI	AMI Meeting Corpus[45]	INRIA	Speech & Text & Speaker ID annotations	Public	WP4
LibriTTS	LibriTTS[46]	INRIA	Speech & transcription	Public	WP2
VCTK	Centre for Speech Technology Voice Cloning Toolkit Corpus[47]	INRIA	Speech & transcription	Public	WP2
VoxForge	VoxForge[48]	INRIA	Speech & transcription	Public	WP3
UCI Housing	scikit learn: the Boston housing prices dataset[49]	INRIA	Sensitive data variables	Public	WP2
Diabetes	scikit learn: the diabetes dataset[50]	INRIA	Diabetes patient record values	Public	WP2
Tripadvisor	UCI: Travel Reviews Data Set[51]	INRIA	User ID & evaluation values	Public	WP2
PTB	Penn Treebank-3[52]	USAAR	Text	Public but paid	WP3
CoNLL-2003	CoNLL-2003 challenge dataset[53]	USAAR	Text & Annotation	Public	WP4
ATIS	Air Travel Information System[54]	USAAR	Text & Annotation	Public	WP4

43 https://github.com/DialRC/LetsGoDataset
44 https://www.robots.ox.ac.uk/~vgg/data/voxceleb/
45 https://groups.inf.ed.ac.uk/ami/corpus/
46 https://research.google/tools/datasets/libri-tts/
47 https://datashare.ed.ac.uk/handle/10283/2950
48 http://www.voxforge.org
49 https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html
50 https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html
51 https://archive.ics.uci.edu/ml/datasets/Travel+Reviews
52 https://catalog.ldc.upenn.edu/LDC99T42
53 https://www.clips.uantwerpen.be/conll2003/ner/
54 https://github.com/Microsoft/CNTK/tree/master/Examples/LanguageUnderstanding/ATIS/Data

Identifier	Title	Partner	Data type	Access	WP
SNIPS	Snips NLU[55]	USAAR	Text & Annotation	Public	WP4
FB en-TOD	Multilingual Task-Oriented Dialog Data[56]	USAAR	Text & Annotation	Public	WP4
MS Taxi	SLT 2018 Special Session - Microsoft Dialogue Challenge (Data: Taxi)[57]	USAAR	Text & Annotation	Public	WP4
MS Restaurant	SLT 2018 Special Session - Microsoft Dialogue Challenge (Data: Restaurant)[58]	USAAR	Text & Annotation	Public	WP4
TwitterAAE	TwitterAAE[59]	USAAR	Text & Annotation	Public	WP4
SADILAR isiXhosa NER	SADILAR isiXhosa NER[60]	USAAR	Text & Annotation	Public	WP4
WikiANN	MMNER[61]	USAAR	Text & Annotation	Public	WP4

Appendix B Dataset forms

To make data FAIR, the data management plan has relied on the collection of data management forms that have been filled out at different stages of the COMPRISE implementation progress. The use of the data management form, given in **Table 6**, has indeed facilitated an automatic integration of COMPRISE data for other purposes allowing interdisciplinary interoperability.

The data management form includes, among others, the data type, the origin of the data, the related work package number, and the format, and in which repository the data has been stored. It also describes the purpose of the data collection in relation with the objectives of the project, as well as the data utility for clarifying to whom the data might be useful.

Table 6: Data management form template

Dataset identifier	The ID allocated using the naming convention outlined in Section 3.1 “Making data findable, including provisions for metadata”
Title of dataset	The title of the dataset which should be easily searchable and findable

[55] https://github.com/snipsco/snips-nlu
[56] https://fb.me/multilingual_task_oriented_data
[57] Taxi dataset from: https://github.com/xiul-msr/e2e_dialog_challenge
[58] Restaurant dataset from: https://github.com/xiul-msr/e2e_dialog_challenge
[59] http://slanglab.cs.umass.edu/TwitterAAE/
[60] https://repo.sadilar.org/handle/20.500.12185/312
[61] https://github.com/afshinrahimi/mmner

Partner	Lead partner responsible for the creation of the dataset
Work package and task(s)	The work package associated to the dataset
Origin	How was the dataset generated?
Data description	A brief description of the dataset
Purpose and relation to COMPRISE objectives	Purpose and relation to the project
Type of data	The type of the dataset, e.g., voice, text, etc.
Utility	To whom the data might be useful?
Expected reuse	Will the data be reused? If yes, how? e.g., research and scientific community, benchmarking, etc.
Type format	This could be DOC, XLSX, PDF, JPEG, TIFF, PPT, etc.
Data size	Size of the dataset
Data capture and processing methods	How was the data collected and processed?
Data repository	Expected repository to be submitted e.g., Institutional/MyBox & Zenodo
DOI	The DOI can be entered once the dataset has been deposited in the repository
Access	Initially how can we have access to the dataset? Will it be open after publication?
Restriction on sharing	Are there any restrictions to share the data or is it publicly available? If restricted, please explain why?
Supporting tools	Software and tools information to use/access the data
Copyright and IP management	In line with the COMPRISE Consortium Agreement
Licensing	e.g., Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)
Quality assurance	What are the quality assurance measures taken? e.g., Recording equipment accuracy tested prior to dataset collection, extensive accuracy measurements conducted prior to the dataset release, etc.
Date of Repository Submission	The date of submission to the repository can be added once it has been submitted
Keywords	The keywords associated with the dataset
Version Number	The version number to keep track of changes to the dataset
Link to metadata file	Link to metadata file

A description of the different datasets that have witnessed revisions by adding an extra set of collected data is presented in the following sub-sections.

Appendix B.1 Drive-Thru beta testers dataset

Table 7: Drive-Thru beta testers dataset

Dataset identifier	COMPRISE_Data01_Drive-Thru_V1.0
Title of dataset	Drive-Thru beta testers dataset
Partner	NETF
Work package and task(s)	WP6
Origin	Proprietary
Data description	This dataset contains all the details needed to correctly make orders online on NETF's Drive-Thru platform. The data consists of customer details (name, address, purchase history, favourite store, etc.) and the list of all goods proposed for sale on the platform
Purpose and relation to COMPRISE objectives	This data has been used to build the Drive-thru voice-based application as described in the objectives of the COMPRISE project. For instance, a Drive-thru customer is able to make online orders on the store using voice features.
Type of data	Textual and relational data
Utility	To NETF only
Expected reuse	No
Type format	SQL and NoSQL databases
Data size	500 GB
Data capture and processing methods	n.a.
Data repository	Private
DOI (if known)	n.a.
Access	Private
Restriction on sharing	No sharing
Supporting tools	Unknown so far
Copyright and IP management	Private
Licensing	Proprietary
Quality assurance	Confidential
Date of Repository Submission	n.a.
Keywords	Drive-thru, e-commerce
Version Number	1.0
Link to metadata file	n.a.

Appendix B.2 Tilde Balss Test Set

Table 8: Tilde Balss Test Set

Dataset identifier	COMPRISE_Data02_TBTS_V1.0
Title of dataset	Tilde Balss Test Set
Partner	TILDE
Work package and task(s)	WP3
Origin	Self-collected
Data description	This data consist of queries, short messages, addresses and other interactions from a two different voice-enabled apps (general voice input and educational), etc. that have been collected in natural environments, i.e., background noise is included in the recordings. Of this Latvian dataset, a subset of clean 1,159 utterances have been translated into English.
Purpose and relation to COMPRISE objectives	This data has been used to study, test, and evaluate Speech-to-Text models for Speech-to-Text Machine Translation of real-world voice-enabled system translation.
Type of data	Audio & Text
Utility	To TILDE only
Expected reuse	Internal reuse only
Type format	Audio wav & text files
Data size	2 GB
Data capture and processing methods	Data was captured on voice-enabled apps in natural use-case environments.
Data repository	Internal
DOI (if known)	n.a.
Access	Private
Restriction on sharing	n.a.
Supporting tools	Any common audio player & text editing tool
Copyright and IP management	Private
Licensing	Proprietary
Quality assurance	Confidential
Date of Repository Submission	n.a.
Keywords	speech-to-text translation
Version Number	V1.0
Link to metadata file	n.a.

Appendix B.3 LibriSpeech dataset

Table 9: LibriSpeech dataset

Dataset identifier	COMPRISE_Data03_LibriSpeech_V1.0
Title of dataset	LibriSpeech
Partner	INRIA
Work package and task(s)	WP2, WP4
Origin	https://www.openslr.org/12/
Data description	LibriSpeech is a corpus of read speech, based on LibriVox's public domain audio books. Its purpose is to enable the training and testing of Speech-to-Text systems. Within COMPRISE, we also used it for the development and evaluation of speech anonymization systems. To do so, we created training/development/test data splits for the evaluation of speaker verification performance.
Purpose and relation to COMPRISE objectives	It is relevant to COMPRISE because speaker identity information is available in addition to ground-truth transcriptions.
Type of data	Read speech of audio books & their transcriptions.
Utility	The corpus has been used for training acoustic and language models for speech-to-text, as well as for developing privacy-driven speech and text transformations.
Expected reuse	Reuse by the scientific community.
Type format	Raw samples (audio), .csv (transcripts)
Data size	150 GB
Data capture and processing methods	Unknown
Data repository	https://doi.org/10.5281/zenodo.5736583
DOI (if known)	10.5281/zenodo.5736583
Access	Publicly available.
Restriction on sharing	In line with the COMPRISE Consortium Agreement
Supporting tools	n.a
Copyright and IP management	Creative Commons Attribution 4.0 International License
Licensing	
Quality assurance	n.a.
Date of Repository Submission	n.a.
Keywords	Read speech, training data
Version Number	1.0
Link to metadata file	n.a.

Appendix B.4 Mozilla Common Voice dataset

Table 10: Mozilla Common Voice dataset

Dataset identifier	COMPRISE_Data04_ CommonVoice_V1.0
Title of dataset	Mozilla Common Voice
Partner	INRIA
Work package and task(s)	WP2
Origin	https://commonvoice.mozilla.org
Data description	The Mozilla Common Voice corpus is a crowdsourced data collection project that aims to diversify the audio data used to train Speech-To-Text models. This diversification is not limited to spoken language, gender, age, race, but also the various accents that are used around the world. Within COMPRISE, we also used the English subset of Mozilla Common Voice for the evaluation of speech anonymization systems. To do so, we created training/development/test data splits for the evaluation of speaker verification performance.
Purpose and relation to COMPRISE objectives	This data provides a rich variety of spoken language data with the associated text. Using the diverse data, COMPRISE is possible to train models that will better understand the utterances of the younger demography, older demography and furthermore, utterances that include regional accents.
Type of data	Read speech & text
Utility	The corpus has been used for training acoustic and language models for speech-to-text, as well as for evaluating privacy-driven speech transformations.
Expected reuse	Reuse by the scientific community.
Type format	Raw samples (audio), .csv (transcripts)
Data size	50 GB
Data capture and processing methods	Crowdsourcing
Data repository	https://doi.org/10.5281/zenodo.5736808
DOI (if known)	10.5281/zenodo.5736808
Access	Publicly available.
Restriction on sharing	In line with the COMPRISE Consortium Agreement
Supporting tools	n.a
Copyright and IP management	Creative Commons Attribution 4.0 International License
Licensing	
Quality assurance	n.a.
Date of Repository Submission	n.a.
Keywords	Read speech, training data

Version Number	1.0
Link to metadata file	n.a.

Appendix B.5 VerbMobil-1 Corpus dataset

Table *11*: VerbMobil-1 Corpus dataset

Dataset identifier	COMPRISE_Data05_VerbMobil-1Corpus_V1.0
Title of dataset	VerbMobil-1 Corpus
Partner	USAAR
Work package and task(s)	WP2, WP3, WP4
Origin	https://www.phonetik.uni-muenchen.de/Bas/BasVM1eng.html
Data description	Recordings of two people trying to reach an agreement for the details of a meeting (time, place, activities, etc.) to be complemented by newly collected privacy-related annotations. Within COMPRISE, we annotated some named entities capturing private information in the conversation like personal name, location, organization, date, and time. We used the annotated corpus to train a named entity recognition model to automatically detect private entities. In total, we annotated 27,000 utterances.
Purpose and relation to COMPRISE objectives	Negotiating times, places and preferences for activities bears a large potential for containing privacy-related information. The corpus thus lends itself naturally to the tasks of privacy-preserving transformations for both speech and text. The latter aspect can be realised because the corpus contains a large amount of both recordings and transcriptions. This is a dialogue corpus, which is another plus for COMPRISE.
Type of data	Audio recordings and their transcriptions.
Utility	The corpus can be used for training acoustic and language models for speech-to-text, dialogue models, as well as for developing privacy-driven speech and text transformations.
Expected reuse	Reuse by the scientific community.
Type format	PhonDat-2 (audio) and ASCII (transcriptions)
Data size	9 GB
Data capture and processing methods	Unknown
Data repository	https://github.com/uds-lsv/privacy-preserving-text-transformer https://doi.org/10.5281/zenodo.5742055
DOI (if known)	10.5281/zenodo.5742055
Access	The corpus is publicly available for a fee
Restriction on sharing	The corpus cannot be shared
Supporting tools	The corpus installation comes with a supporting code to process the data

Copyright and IP management	The corpus is distributed by the Bavarian Archive for Speech Signals (BAS)
Licensing	Proprietary license
Quality assurance	n.a.
Date of Repository Submission	n.a.
Keywords	Spoken dialogue, negotiation.
Version Number	1.0
Link to metadata file	n.a.

Appendix B.6 Yelp dataset

Table 12: Yelp dataset

Dataset identifier	COMPRISE_Data06_YELP_V1.0
Title of dataset	YELP Style Transfer dataset with Gender and Sentiment Annotations
Partner	USAAR
Work package and task(s)	WP2
Origin	https://github.com/uds-lsv/author-profiling-prevention-BT
Data description	The dataset contains Yelp reviews, each review has both gender and sentiment annotation. The texts are obtained from two datasets (YelpGender & Yelp Sentiment) but from the same source. By automatically comparing each review in the test set of YelpGender with the YelpSentiment Dev and Test sets, we created a new Dev set and Test set with 4,000 reviews, each with both gender and sentiment information. This can be used for future research to evaluate the utility of the Yelp Gender dataset.
Purpose and relation to COMPRISE objectives	The purpose of the dataset is to evaluate the utility of style transfer models on downstream NLP performance.
Type of data	Textual data
Utility	The dataset was used to evaluate the performance of style transfer models for anonymizing gender attributes.
Expected reuse	Reuse by the scientific community
Type format	TSV
Data size	355 kB
Data capture and processing methods	Unknown
Data repository	https://github.com/uds-lsv/author-profiling-prevention-BT
DOI (if known)	Unknown
Access	Publicly available
Restriction on sharing	No restriction
Supporting tools	n.a

Copyright and IP management	In line with the COMPRISE Consortium Agreement
Licensing	MIT License
Quality assurance	n.a
Date of Repository Submission	n.a
Keywords	style transfer, Yelp Gender, Yelp Sentiment
Version Number	n.a
Link to metadata file	n.a

Appendix B.7 Hausa VOA NER dataset

Table 13: Hausa VOA NER dataset

Dataset identifier	COMPRISE_Data08_H-VOA-NER_V1.0
Title of dataset	VOA Named Entity Recognition dataset for Hausa
Partner	USAAR
Work package and task(s)	WP4
Origin	https://github.com/uds-lsv/transfer-distant-transformer-african/tree/master/data/hausa_ner
Data description	A named entity recognition dataset (a subset of the News dataset). This dataset that consists of 1,450 sentences has been annotated for the following four entity types: Personal name, Locations, Organizations, and dates.
Purpose and relation to COMPRISE objectives	The purpose of the dataset is to evaluate the performance of different weakly supervised learning techniques for named entity recognition in low-resourced languages.
Type of data	Textual data with named entity annotations
Utility	The dataset has been used to train weakly supervised learning models for named entity recognition.
Expected reuse	Reuse by the scientific community
Type format	TXT (CoNLL data Format)
Data size	325 kB
Data capture and processing methods	The text was collected from Voice of America's news website for Hausa and annotated by two native speakers.
Data repository	Institutional
DOI (if known)	Unknown
Access	Publicly available
Restriction on sharing	No restriction
Supporting tools	n.a
Copyright and IP management	In line with the COMPRISE Consortium Agreement
Licensing	Creative Commons Attribution 4.0 International (CC BY 4.0)

Quality assurance	n.a
Date of Repository Submission	n.a
Keywords	named entity recognition, NER, VOA, Hausa
Version Number	n.a
Link to metadata file	n.a

Appendix B.8 Hausa VOA Topics dataset

Table 14: Hausa VOA Topics dataset

Dataset identifier	COMPRISE_Data09_H-VOA-Topics_V1.0
Title of dataset	Hausa VOA News Topics Classification Dataset
Partner	USAAR
Work package and task(s)	WP4
Origin	https://github.com/uds-lsv/transfer-distant-transformer-african/tree/master/data/hausa_newsclass
Data description	A text classification dataset to categorize news headlines into topics. This dataset that contains 2,917 sentences has been annotated with the five following entity types: Nigeria, Africa, World, Health, and Politics.
Purpose and relation to COMPRISE objectives	The purpose of the dataset is to evaluate the performance of different weakly supervised learning techniques for text classification in low-resourced languages.
Type of data	Textual data with annotations
Utility	The dataset has been used to train weakly supervised learning models for text classification.
Expected reuse	Reuse by the scientific community
Type format	TSV
Data size	196 kB
Data capture and processing methods	The text was collected from Voice of America's news website for Hausa and annotated by two native speakers.
Data repository	Institutional
DOI (if known)	Unknown
Access	Publicly available
Restriction on sharing	No restriction
Supporting tools	n.a
Copyright and IP management	In line with the COMPRISE Consortium Agreement
Licensing	Creative Commons Attribution 4.0 International (CC BY 4.0)
Quality assurance	n.a
Date of Repository Submission	n.a

Keywords	news topics, text classification, BBC, Hausa
Version Number	n.a
Link to metadata file	n.a

Appendix B.9 MENYO-20K dataset

Table 15: MENYO-20K dataset

Dataset identifier	COMPRISE_Data12_MENYO-20K_V1.0
Title of dataset	MENYO-20k: A Multi-domain English-Yorùbá Corpus for Machine Translation and Domain Adaptation
Partner	USAAR
Work package and task(s)	WP3
Origin	https://github.com/uds-lsv/menyo-20k_MT
Data description	A multi-domain parallel corpus for English-Yoruba language pair that can be used to benchmark machine translation systems. The dataset has 20,100 parallel sentences split into 10,070 train, 3,397 dev, and 6,633 test sentences.
Purpose and relation to COMPRISE objectives	The purpose of the dataset is to evaluate the performance of machine translation models on low-resource languages
Type of data	Textual data
Utility	The dataset has been used to develop machine translation models for low-resource languages.
Expected reuse	Reuse by the scientific community
Type format	TSV
Data size	266 kB
Data capture and processing methods	The MENYO-20k corpus is obtained from several domains such as news articles, ted talks, movie transcripts, etc.
Data repository	Institutional
DOI (if known)	Unknown
Access	Publicly available
Restriction on sharing	No restriction
Supporting tools	n.a
Copyright and IP management	In line with the COMPRISE Consortium Agreement
Licensing	Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
Quality assurance	n.a
Date of Repository Submission	n.a
Keywords	multi-domain, machine translation, MENYO-20k, Yoruba
Version Number	n.a

Link to metadata file	n.a

Appendix B.10 MasakhaNER dataset

Table 16: MasakhaNER dataset

Dataset identifier	COMPRISE_Data13_M-NER_V1.0
Title of dataset	MasakhaNER: Named Entity Recognition for African Languages
Partner	USAAR
Work package and task(s)	WP4
Origin	https://github.com/masakhane-io/masakhane-ner
Data description	A named entity recognition (NER) dataset (a subset of the News dataset). It has four entity types: Personal names, Locations, Organizations, and Dates. The dataset supports ten African languages: Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Nigerian-Pidgin, Swahili, Wolof, and Yoruba.
Purpose and relation to COMPRISE objectives	The purpose of the dataset is to evaluate the performance of different weakly supervised learning techniques for NER
Type of data	Textual data with named entity annotations
Utility	The dataset has been used to train weakly supervised learning models for named entity recognition.
Expected reuse	Reuse by the scientific community
Type format	TXT (CoNLL data Format)
Data size	5.4 MB
Data capture and processing methods	Unknown
Data repository	Institutional
DOI (if known)	Unknown
Access	Publicly available
Restriction on sharing	No restriction
Supporting tools	n.a
Copyright and IP management	In line with the COMPRISE Consortium Agreement
Licensing	Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
Quality assurance	n.a
Date of Repository Submission	n.a
Keywords	named entity recognition, NER, MasakhaNER
Version Number	n.a
Link to metadata file	n.a

Appendix B.11 Yoruba GV NER dataset

Table 17: Yoruba GV NER Data

Dataset identifier	COMPRISE_Data14_Y-GV-NER_V1.0
Title of dataset	Global Voices Named Entity Recognition dataset for Yoruba
Partner	USAAR
Work package and task(s)	WP4
Origin	https://github.com/ajesujoba/YorubaTwi-Embedding/tree/master/Yoruba/Yoruba-NER
Data description	A named entity recognition dataset (a subset of the News dataset). It has four entity types: personal name, location, organization and date. It contains 1,168 sentences.
Purpose and relation to COMPRISE objectives	The purpose of the dataset is to evaluate the performance of different weakly supervised learning techniques for named entity recognition in low-resourced languages.
Type of data	Textual data with named entity annotations
Utility	The dataset has been used to train weakly supervised learning models for named entity recognition.
Expected reuse	Reuse by the scientific community
Type format	TXT (CoNLL data Format)
Data size	254 kB
Data capture and processing methods	The text was collected from Global Voices news website for Yoruba and annotated by two native speakers.
Data repository	Institutional
DOI (if known)	Unknown
Access	Publicly available
Restriction on sharing	No restriction
Supporting tools	n.a
Copyright and IP management	In line with the COMPRISE Consortium Agreement
Licensing	Creative Commons Attribution 3.0 International (CC BY 3.0)
Quality assurance	n.a
Date of Repository Submission	n.a
Keywords	named entity recognition, NER, Global Voices, Yoruba
Version Number	n.a
Link to metadata file	n.a

Appendix B.12 Yoruba BBC Topics dataset

Table 18: Yoruba BBC Topics dataset

Dataset identifier	COMPRISE_Data15_Y-BBCTopics_V1.0
Title of dataset	Yoruba BBC News Topics Classification Dataset
Partner	USAAR
Work package and task(s)	WP4
Origin	https://github.com/uds-lsv/transfer-distant-transformer-african/tree/master/data/yoruba_newsclass
Data description	A text classification dataset to categorize news headlines into topics. It has seven categories: Sports, Entertainment, Nigeria, Africa, World, Health, and Politics. It contains 1,908 sentences.
Purpose and relation to COMPRISE objectives	The purpose of the dataset is to evaluate the performance of different weakly supervised learning techniques for text classification in low-resourced languages.
Type of data	Textual data with annotations
Utility	The dataset has been used to train weakly supervised learning models for text classification.
Expected reuse	Reuse by the scientific community
Type format	TSV
Data size	266 kB
Data capture and processing methods	The text was collected from the BBC news website for Yoruba and annotated by two native speakers.
Data repository	Institutional
DOI (if known)	Unknown
Access	Publicly available
Restriction on sharing	No restriction
Supporting tools	n.a
Copyright and IP management	In line with the COMPRISE Consortium Agreement
Licensing	Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
Quality assurance	n.a
Date of Repository Submission	n.a
Keywords	news topics, text classification, BBC, Yoruba
Version Number	n.a
Link to metadata file	n.a

Appendix B.13 Notes app dataset

Table 19: Notes App dataset

Dataset identifier	COMPRISE_Data16_Notes-ED_V1.0
Title of dataset	Notes app Dataset
Partner	ASCO
Work package and task(s)	WP6
Origin	Self-collected
Data description	This dataset contains speech audio files, Speech-to-Text output, Text Transformation results of the user test of the COMPRISE Notes app. The subjects' dialogue was elicited by asking categorical questions, but where free to utter phrases within said category (e.g., Make a Xmas shopping list).
Purpose and relation to COMPRISE objectives	This dataset was created to evaluate the COMPRISE components when used on real users of a note taking app.
Type of data	Audio and text
Utility	For internal use only
Expected reuse	Internal reuse only
Type format	Audio, text, survey results and annotations
Data size	30 MB
Data capture and processing methods	The audio and text data were captured using the COMPRISE demonstrator app, Notes, and the annotations and transcriptions have been verified manually.
Data repository	Private
DOI (if known)	n.a.
Access	Internal
Restriction on sharing	Confidential
Supporting tools	n.a.
Copyright and IP management	Private
Licensing	Proprietary
Quality assurance	Private
Date of Repository Submission	n.a.
Keywords	User interaction data
Version Number	1.0
Link to metadata file	n.a.

Appendix B.14 Shoplay dataset

Table 20: Shoplay dataset

Dataset identifier	COMPRISE_Data17_ShoplayED_V1.0
Title of dataset	Shoplay dataset
Partner	NETF
Work package and task(s)	WP6
Origin	Self-collected
Data description	This dataset contains audio and text files of 82 evaluation sessions recorded during the Shoplay app evaluation process on users of the app. The data is of subjects using voice as a means to navigate grocery items that they wish to buy, or learn the names of various grocery items in different languages.
Purpose and relation to COMPRISE objectives	This data is used to assess the quality of the output from the COMPRISE components when used in the domain of e-commerce.
Type of data	Audio and textual
Utility	To NETF only
Expected reuse	No
Type format	Audio and text outputs
Data size	11 MB
Data capture and processing methods	Data was captured using the COMPRISE demonstrator app, Shoplay, and processed internally.
Data repository	Private
DOI (if known)	n.a.
Access	Private
Restriction on sharing	No sharing
Supporting tools	Unknown so far
Copyright and IP management	Private
Licensing	Proprietary
Quality assurance	Confidential
Date of Repository Submission	n.a.
Keywords	Shoplay, e-commerce
Version Number	1.0
Link to metadata file	n.a.

Appendix B.15 Hospital Concierge dataset

Table 21: Hospital Concierge dataset

Dataset identifier	COMPRISE_Data18_HCED_V1.0
Title of dataset	Hospital Concierge dataset
Partner	TILDE
Work package and task(s)	WP6
Origin	Self-collected
Data description	This data has been collected as part of the testing and evaluation of the Hospital Concierge system that is under proto-type use at a Latvian hospital. The data consists of audio recordings, privacy converted audio files, Speech-to-Text outputs, and transcription of the audio files.
Purpose and relation to COMPRISE objectives	The prototype that collected this data is placed at the entrance of a hospital, and therefore the identity of the users and the situations they might utter to the system is highly confidential. Furthermore, the user of this system may not always be in a stable mindset, which create domain specific adaptation need. The data collected from this multi-lingual system will therefore provide much needed information and opportunities to demonstrate the possibilities of training domain specific models without the vast requirement of huge, annotated datasets.
Type of data	Audio & Excel files: • Speech-to-Text/reference transcriptions, • text input from users, • detected/reference intents for both voice and text input.
Utility	To TILDE only
Expected reuse	Internal reuse only
Type format	Audio wav & text files
Data size	76 MB
Data capture and processing methods	Data was collected a hospital located in Riga and processed internally by TILDE
Data repository	Internal
DOI (if known)	n.a.
Access	Private
Restriction on sharing	n.a.
Supporting tools	Any common audio player & text editing tool
Copyright and IP management	Private
Licensing	Proprietary
Quality assurance	Confidential

Date of Repository Submission	n.a.
Keywords	Medical dialogues
Version Number	V1.0
Link to metadata file	n.a.

Appendix B.16 Doctor's Assistant dataset

Table 22: Doctor's Assistant dataset

Dataset identifier	COMPRISE_Data19_DA-ED_V1.0
Title of dataset	Doctor's Assistant dataset
Partner	TILDE
Work package and task(s)	WP6
Origin	Self-collected
Data description	This data has been collected as part of the testing and evaluation of the Doctor's Assistant app that is under development for future usage in Latvian hospitals. The data consists of audio recordings, privacy converted audio files, Speech-to-Text outputs, and transcription of the audio files.
Purpose and relation to COMPRISE objectives	This dataset is at the core of the COMPRISE values since the app is planned to be used in an extremely privacy sensitive environment. Therefore, this data is used to find and eliminate as far as possible the weakness that may be present in relation to privacy and also create Speech-to-Text and Natural Language Understanding models that will be able to understand to a greater accuracy the domain specific utterances of real working medical doctors.
Type of data	Audio & Excel files: • Speech-to-Text/reference transcriptions • text input from users • detected/reference intents for both voice and text input • speaker labels for voice inputs
Utility	To TILDE only
Expected reuse	Internal reuse only
Type format	Audio wav & text files
Data size	218 MB
Data capture and processing methods	Data was collected a hospital located in Riga and processed internally by TILDE
Data repository	Internal
DOI (if known)	n.a.
Access	Private
Restriction on sharing	n.a.
Supporting tools	Any common audio player & text editing tool

Copyright and IP management	Private
Licensing	Proprietary
Quality assurance	Confidential
Date of Repository Submission	n.a.
Keywords	Medical dialogues
Version Number	V1.0
Link to metadata file	n.a.

Appendix C Metadata file template

Adding to the Zenodo metadata schema (discussed in Section 3.2 “Making data openly accessible”), the uploaded datasets have been accompanied by a metadata file containing the information provided in the data management form in Appendix B “Dataset forms”.

Table 23: Metadata file description

This metadata file was generated on <insert_date> by <insert_name>

GENERAL INFORMATION

---

1. Title of dataset:

2. Dataset identifier in the repository:

3. Dataset DOI:

4. Responsible partner:

5. Author(s) information:

Contact Information 1

Role:

Name:

Email:

Organisation:

Contact Information 2

Role:

Name:

Email:

Organisation:

6. Period of data collection:

7. Geographic location of data collection:

8. The title of project and Funding sources that supported the collection of the data:

COMPRISE has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 825081.

SHARING/ACCESS INFORMATION

---

1. License/access restrictions placed on the data:

2. Link to data repository:

3. Was data derived from another source?

If yes, list source(s):

DATASET OVERVIEW

---

1. Sub-datasets included:

2. Status of the documented data? – “complete”, “in progress”, or “planned”

Are there any plans to update the data?

3. Keywords:

METHODOLOGICAL INFORMATION

---

1. Methods used for experimental design and data collection:

2. Methods used for data processing:

3. Instruments and software used in data collection and processing:

5. Experimental conditions:

6. Quality-assurance procedures performed on the data:

7. Dataset benefit: