COMPRISE

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 Nº1.3: Initial Data Management Plan

Lead partner: INRIA

Version Nº: 1.0

Date: 22/05/2019

Document information	
Deliverable Nº and title:	D1.3 – Initial Data Management Plan
Version Nº:	1.0
Lead beneficiary:	INRIA
Author(s):	Zaineb Chelly Dagdia (INRIA), Emmanuel Vincent (INRIA)
Reviewers:	Gerrit Klasen (ASCO), Álvaro Moretón (ROOT)
Submission date:	22/05/2019
Due date:	31/05/2019
Type[1]:	ORDP
Dissemination level[2]:	PU

Document history			
Date	Version	Author(s)	Comments
09/04/2019	0.1	Zaineb Chelly	Draft deliverable
22/04/2019	0.2	Zaineb Chelly	Initial version
17/05/2019	1.0	Zaineb Chelly & 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 aims at planning, organizing and controlling all activities and tasks so that the project successfully runs throughout all its stages. The activities involved in WP1 include 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 will contribute 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 are supporting this effort.

So far, two WP1 deliverables were submitted at M2 and M3. The first deliverable describes the "Project quality plan and private web platform" (D1.1 − M2) and the second deliverable explains the "Detailed work plan" (D1.2 − M3). In the current document, the third deliverable of WP1 entitled "Initial Data Management Plan" (D1.3) is presented.

D1.3 explains how the COMPRISE consortium is planning to manage the data gathered in the course of the project. This includes planning how data will be collected, used, managed, stored, sustainably archived, and disseminated.

This document is a living document that will be further updated as the implementation of the project progresses and when significant changes occur. A final version of the DMP will be submitted to the European Commission by M36.

Table of contents

1. Introduction

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

This document presents a preliminary version of the DMP submitted to the European Commission on M6 of COMPRISE. It is a living document that will be further updated throughout the project lifecycle whenever significant changes arise.

This document, which will be publicly disseminated, will serve as a guidance to all COMPRISE members on how to manage data, and to agree on the opted management policy plan. This will enable the consortium members to have a common understanding and, where possible, shared practices.

D1.3 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 Center (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 will be managed over the course of COMPRISE. Section 3 details FAIR data explaining 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.

[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. Data summary

This section provides a preliminary 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.

2.1 Data purpose and data utility

As previously defined in D1.2 – "Detailed work plan" (submitted to the European Commission on February 28, 2019 − Confidential), within WP6 – "Evaluation and demonstration for practical use cases", common research datasets will be defined and/or collected 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 will be gathered for the development and evaluation of the three demonstrators, i.e., Consumer application development (T6.2), e-commerce (T6.3), and e-health (T6.4), corresponding to real industrial use cases.

The definition of common datasets will allow the scientific validation of the project's components and tools in combination with each other, and their evaluation within the demonstrators (T6.1). 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 will be assessed and acquired for the specific following purposes:

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

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

Apart from being used by the COMPRISE partners for the above purposes, the defined datasets may be useful to several external parties and stakeholders who have been identified in D7.1 − "Dissemination and communication action plan" (submitted to the European Commission on February 28, 2019 − Public), and in D1.2 − "Detailed work plan" (submitted to the European Commission on February 28, 2019 − Confidential). These external parties mainly include:

- the research and scientific community in the fields of machine learning, speech and language processing, privacy, etc., who will be able to use the datasets as benchmarks for experimental studies,
- end-users who will be able to have an increased range of in-app functionalities,
- software developers in the field of voice-enabled technologies who will be able to have a data foundation for their own solutions.

2.2 COMPRISE datasets

For the current preliminary version of the DMP, and with respect to the data purposes previously defined, the consortium members identified several datasets to be used and collected for each specified aim. Some of the datasets defined for research purposes will be used in their current state (see Appendix A), while some others will be enriched with additional collected data. As for demonstrators, some existing datasets will be enriched

with extra data, while others will be specifically built from scratch for each demonstrator's industrial use-case.

In the following, we only discuss the existing datasets that are expected to be modified and/or enriched within COMPRISE. A summary of these is given in Table 1, whereas a full detailed description is provided in Appendix B. A detailed description of the unmodified datasets falls out of the scope of this document, while datasets to be collected from scratch at a later stage will be described in a future version of this document.

Table 1: Preliminary list of COMPRISE datasets.

Datasets for research purposes					
Identifier	**Title**	**Partner**	**Data type**	**Access**	**WP & Task**
COMPRISE_Data1_VerbMobil-1Corpus_V1.0	VerbMobil-1 Corpus	USAAR	Audio & Transcripts	Public but paid	WP3, T3.2
COMPRISE_Data2_Lets_Go!_V1.0	Let's-Go! dataset	USAAR	Audio & Transcripts	Public	WP3, T3.2
Datasets for demonstrators					
Identifier	**Title**	**Partner**	**Data type**	**Access**	**WP & Task**
COMPRISE_Data3_Drive-Thru_V1.0	Drive-Thru beta testers dataset	NETF	Text and relational data	Confidential	WP6 T6.3
COMPRISE_Data4_RecipeBox_V1.0	RecipeBox dataset	ASCO	JSON	Public	WP6 T6.2

The two datasets identified for research purposes are considered as benchmark datasets and already publicly available. They will be reused and modified to achieve the project goals. This is also the case for some of the demonstrators' datasets, e.g., "RecipeBox". Additionally, the datasets collected as a complement to the public datasets are expected to have long-term value and utility. Apart from being useful to the validation of COMPRISE research, the collected data may be reused for other research studies, benchmarking, reproducible research, scientific impact maximization, and within exploitation activities planned by the partners.

More precisely, the two identified research datasets that are intended to be revised, i.e., VerbMobil-1 and LetsGo!, consist of general conversations and currently do not contain explicitly privacy-related annotations. However, for the evaluation of the privacy-driven speech and text transformations in COMPRISE, such ground-truth data is required. Additionally, for approaches based on supervised machine learning, data annotations are required for training the classifiers. Existing metadata on speaker's gender, age, and possibly profession is available and will be used to this end, but other potentially private information is currently not covered. Therefore, USAAR will carry out additional annotations on suitable subsets of these corpora. A first annotation round will focus on the geographical location of users at the time of the recording as well as their stated preferences regarding locations for appointments. This annotation will serve as a general testbed for both the annotation process as well as the development of initial transformation models. Annotations performed locally by USAAR staff as well as internet-based annotations though crowdsourcing platforms will both be considered. All annotations will be published in open access in forms similar to the COMPRISE data collections.

The RecipeBox demonstrator dataset, which is also intended to be revised, provides multiple cooking recipes and descriptions on how to prepare these. It is assumed that translated versions of these recipes and/or extended instructions for the users will have to be added as part of the development process for the corresponding demonstrator. This additional data will also be published in open access.

Conversely, the data to be collected from beta-testers or end-users for the evaluation of the demonstrators will remain confidential, due to its increased realism possibly implying real private information and in order to preserve the demonstrators' competitive advantage until the eventual public release of the corresponding product. Consequently, this data will neither be made publicly available nor reused in research studies outside the COMPRISE environment. Yet, they will play an important role within COMPRISE to validate and improve the desired functionalities. The expected utility for each defined dataset is detailed in Appendix B.

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

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

As previously mentioned, a majority of the used datasets will be openly accessible, while some of the collected datasets will be confidential. For all openly accessible datasets, and to maximize their interoperability, management, and reuse, the consortium members agreed to use, whenever possible, formats that are non-proprietary, unencrypted, uncompressed and in common usage by the research community. To this end, the consortium members have agreed to follow whenever possible the indications of the UK Data Archive[11], as recommended by OpenAIRE[12], 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 not present in 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	Hypertext Mark-up Language (.html) Widely-used formats: MS Word (.doc/.docx)

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

	according to an appropriate Document Type Definition (DTD) or schema	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)
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 the COMPRISE implementation and before the dissemination of the results, the collected datasets will be securely stored in the project's internal communication tool named MyBox which was described in Section 4.1 of "D1.1 − Project quality plan and private web platform" (submitted to the European Commission on January 31, 2019 − Confidential). MyBox allows a storage limit of 10 GB per member. Once the project results are disseminated, the collected data tied to the public datasets will be deposited on the Zenodo[13] platform where a project page (community) has been set up and linked to the COMPRISE website: https://zenodo.org/communities/comprise/. Partners wishing to deposit new datasets can access the upload URL https://zenodo.org/deposit/new?c=comprise that will automatically add new records to the community. Zenodo allows a maximum upload limit of 50 GB. All datasets that will be made publicly available are expected to fit into this limit.

On the other hand, the confidential demonstrator datasets will have an estimated size of several tens of GB. These will be stored on each responsible partner's storage facilities (see Section 5).

Should the storage limit be too small for MyBox and/or Zenodo, the consortium members will contact MyBox/Zenodo in order to increase the storage limit to larger datasets. In any case, the same data (or part of it) will never be uploaded in different entries. This is to avoid the creation of multiple persistent identifiers which makes references and citation difficult.

3. FAIR data

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

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) will be assigned to each uploaded open dataset. This DOI, which is automatically provided by Zenodo to all publicly

[13] http://about.zenodo.org/

available uploads, can be used in any relevant publications to direct readers to the underlying dataset.

So far, the process of collecting open datasets for COMPRISE is in progress. Hence, in this preliminary version of the DMP, there is no defined form yet that provides a specific DOI. This information will be added together with the persistent link in future updates of the current initial version of the DMP, i.e., Deliverable D1.3. This will result in several internal revisions and to a final formal DMP version that will be submitted to the European Commission on M36 (November 2021, D1.6).

To further emphasise how to make data findable, D1.3 as well as the upcoming updated DMP, will rely on the collection of data management forms that will be filled out at different stages of the COMPRISE implementation progress. These forms will be filled by the project partners who are responsible for collecting data. The designed data management forms are only accessible via the project's internal communication tool named Partage while the data itself will be 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. Both of these communication tools, i.e., Partage and MyBox were described in Section 4.1 of "D1.1 − Project quality plan and private web platform" (submitted to the European Commission on January 31, 2019 − Confidential).

The COMPRISE DMP versions will support research reproducibility and trustworthiness, as they will emphasise the need to accurately cite and identify the exact version of the dataset used as a research input, and that underpins research findings. This is crucial because during the project's implementation progress a new version of a dataset may be created when an existing dataset can be reprocessed, corrected or appended with additional data. In this concern, the project's DMP is based on a consistent version numbering scheme[14] that will enable the consortium members to track changes within a 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 of the used data will be identified by a versioning indicator and a history. The opted versioning scheme is based on a two-part numbering format[15]: 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 may reflect 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 may reflect 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.

Only minor changes are expected along the project's implementation although major revisions are possible beyond the end of COMPRISE. This versioning scheme will

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

guarantee specificity and verifiability and will enable each version of a given data collection to be uniquely referenced.

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

The above schema will be used in COMPRISE to establish a naming convention for the project datasets. This will comprise the following items:

- a prefix "COMPRISE" indicating that the dataset was prepared in the course of COMPRISE,
- "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 will be identified as: "COMPRISE_Data1_VerbMobil-1Corpus_V1.0".

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

Every dataset will be annotated with metadata that includes the identifier of the dataset. As COMPRISE relies on the Zenodo repository, it will benefit from the Zenodo JSON metadata schema[16] 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 will be accompanied by a metadata file (see Appendix C) containing the information provided in the data management form given in Appendix B. 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 will be both updated and uploaded to the Zenodo repository.

3.2 Making data openly accessible

COMPRISE will collect a variety of datasets, which have different natures and access privileges. These different access privileges are described in detail in Appendix B and are reviewed here in a concise manner. Appendix B also provides a detailed description

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

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) and potential limitations due to the protection of personal data (see Section 6), different levels of confidentiality are considered within the project consortium:

- **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 is applied to most COMPRISE datasets.

Table 1 in Section 2.1 highlights the publicly available research datasets to be used in COMPRISE that will witness a revision 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 will also fulfil their usefulness for scientific or other public purposes. Regarding the paid datasets, these will be 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 will fulfil the own 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 will be made publicly available, after negotiation with the copyright holder whenever required. These cases can also be in line with some of the demonstrator datasets.

The datasets gathered as part of the demonstrator evaluation process will be 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 will not impact the eventual dissemination of the project outcomes in terms of voice interaction technology. The consortium member in charge will archive the data as set out in Section 5. Like all confidential data in COMPRISE, its preservation and maintenance during and after the project will be handled by the data owners.

As previously mentioned, to facilitate deposit, update, and management, the collected public data will be made available via the Zenodo COMPRISE created community. As for publications, these will be made as open access via the local host institutions repositories, e.g., the European Public Repository HAL[17] for Inria.

In line with Zenodo policies, when uploading public datasets, the consortium members will have to select an option among the following:

- **Open Access.** This is the highly recommended option which provides free access and rights to data. This is mandatory under the H2020 programme[18].
- **Embargoed Access.** This option concerns data underpinning publication. Data will indeed be deposited as soon as possible but open access will be provided

[17] https://hal.inria.fr/

[18] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

only once the data has been published in a scientific paper to preserve the authorship of all authors involved. In such case, information about data will be published and details of when the data will become available will be included in the metadata.

- **Restricted Access.** This option, although not recommended, will be adopted for those data with an access that should be monitored and approved by the depositor if certain requirements to be defined are met.
- **Closed Access.** This option concerns private (but not confidential) datasets.

Although the embargoed or closed access options provided by Zenodo could be a valid choice, the consortium agrees that the confidential datasets to be collected will not be deposited to avoid compromising their protection or commercialisation prospects. This is based on the following Zenodo security statement: *"The files are however stored unencrypted and may be viewed by Zenodo operational staff under specific conditions. This means that "closed access" on Zenodo is not suitable for secret or confidential data."*[19]

Visibility and access to publicly shared datasets will be facilitated by Zenodo metadata and search facility as well as to the automatic link to both OpenAIRE and to the CORDIS project page[20].

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).

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 plan to use SoftwareHeritage[21]. Only open source software is planned to be publicly shared within this platform. On the other side, software which is not publicly released will be either in GitLab[22], 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 aims to collect and document 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 will be fully documented via the data management forms (see Appendix B) and accompanied with detailed metadata supported by a set of select keywords (see Appendix C). This is to facilitate an automatic integration of COMPRISE data for other purposes allowing interdisciplinary interoperability. All data will be provided in generally used extensions, as described in Section 2.3, adopting well established formats whenever possible which will also facilitate its reuse by other parties.

Standard vocabulary will be used for all data types present in the dataset to allow interdisciplinary interoperability. In addition, whenever required, the documentation will

[19] http://about.zenodo.org/infrastructure/
[20] https://cordis.europa.eu/project/rcn/218720/factsheet/en
[21] https://www.softwareheritage.org/
[22] https://about.gitlab.com/

include 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 will be made openly available, once ready for dissemination. To allow the widest possible reuse, the consortium will attach a specific license to every deposited dataset. This will allow indeed 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)[23] license will be used, in order to allow third parties to share and adapt data with no restrictions if attribution is provided.

In case the partner would like to further limit access to the uploaded data, an alternative license will be selected among the following options offered by Zenodo[24]:

- **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.
- **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 will be stored in Zenodo, once ready for dissemination and at the latest upon publication of the related scientific publication, 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 aspects.

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)[25], which currently has an experimental programme defined at least

[23] https://creativecommons.org/licenses/by/4.0/
[24] https://zenodo.org/record/1488616#.XLoeYOgzY2w
[25] http://information-technology.web.cern.ch/about/computer-centre

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 additional costs foreseen for data management are indeed related to:

- the working time needed to set up and perform the data collection and analysis activities,
- the working time to set up and maintain local and shared data collection devices/servers,
- the working time needed to write documentation, metadata, etc.

Dedicated financial resources have been already allocated in each partner budget and specified for some datasets as defined in the Grant Agreement.

The project coordinator is in charge of 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, is the responsibility of the partner that gathers the data in its related work package. Quality control of these data is the responsibility of the relevant work package leader, supported by the Project Coordinator. Each partner should respect the policies set out in this DMP.

Publications featuring the data will be produced in the project and will be made available in 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. Possible costs related to open access will be claimed as part of the Horizon 2020 grant.

Finally, in line with the Consortium Agreement (Article 8.6.2.1), each partner should give at least 21 days prior notice to the other partners before disseminating/publishing data.

5. Data security

During the course of COMPRISE, the collected data related to the public datasets will be stored in MyBox while the collected data tied to the confidential datasets will be stored in the responsible partner's storage facilities, as detailed in Table 3.

Table 3: Data storage description.

Partner institution/organisation	Data storage
INRIA	MyBox[26]: 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.

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

	This makes it possible to share the data with different subgroups of partners, as appropriate.
USAAR	Data will be 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 will be 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 will be 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 will be 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.

As soon as the publication targets are achieved, the public collected datasets will be deposited on Zenodo as previously described. More precisely, open data security will be addressed by taking advantage of Zenodo's services of secure storage, backup and preservation and protected transfer mechanisms.

Regarding the collected data tied to the confidential datasets, different approaches will be used by each data-owner's organisation, but common rules apply. As presented in Table 3, data will be 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., will be granted to users and authorised computers by relevant staff, according to a well-defined protocol. 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 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 are to comply 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 will comply with the ethical principles as set out in the European Code of Conduct for

Research Integrity[27]; these include, in particular, avoiding fabrication, falsification, plagiarism or other research misconduct.

The COMPRISE project will focus 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 shall be used, not only during the operating phase, but also during the training phase.

In this regard, the COMPRISE project will address 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, will delete as much personal information as possible from the users' speech thanks to privacy-driven transformations.

Other important objective of the COMPRISE project will be the creation of demonstrators which will cover 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 will be 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 will be obtained in order 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 may be patients in partnering hospitals. Additional measures will be taken in this case, in collaboration with doctors at the hospital such as the provision of an incidental findings policy.

Further processing of previously collected personal data might be needed as well but, as previously elaborated, the corresponding informed consent will be obtained and all the information regarding such secondary use will be provided, if necessary.

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

All COMPRISE activities raising ethical issues must comply 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 will also comply with additional measures set out in Deliverable "D5.1 −Data protection and GDPR requirements" (Submitted to the European Commission on May 31, 2019 − Confidential).

7. Conclusion

Deliverable D1.3 presents a preliminary version of the data management plan. Throughout this document, a preliminary list of datasets to be collected or modified is presented. These are accompanied by their technical description. A discussion on 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

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

recovery as well as secure storage, together with a description of some other identified ethical aspects were also given.

D1.3 is a living document that will be further updated through the project lifecycle whenever significant changes arise. This will result in several internal revisions and to a final formal DMP version that will be submitted to the European Commission on M36 (November 2021, D1.6). This will be achieved by adding specific details making the information available on a finer level of granularity, and by including needed adjustments and rectifications as the implementation of the project progresses and when significant changes occur.

8. Appendices

8.1 Appendix A. Unrevised datasets

Some of the datasets defined for research purposes will be used in their current state, and hence they will not witness any revisions, e.g. no additional data will be collected to complement the datasets. Table 4 presents a preliminary version of the defined used datasets that will not be revised. Most of these datasets are public.

Table 4: Preliminary list of the unrevised datasets

Identifier	Title	Partner	Data type	Access	WP
IARPA Babel LT	IARPA Babel Lithuanian Language Pack	TILDE	Speech & Text	Public but paid	WP3, WP6
LVDistated	Latvian Dictated Speech corpus	TILDE	Speech & Text	Paid	WP3, WP6
LVASR	Latvian Speech Recognition Corpus	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	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
Portuguese SpeechDat(II)	Portuguese SpeechDat(II) FDB-4000	NETF	Speech & Text & Lexicon	Public but paid	WP6
Librispeech	Librispeech ASR corpus	INRIA	Speech & Text & Speaker ID annotations	Public	WP2, WP3

8.2 Appendix B. Dataset forms

To make data FAIR, the data management plan will rely on the collection of data management forms that will be filled out at different stages of the COMPRISE implementation progress. The use of a data management form, given in Table 5, will indeed facilitate 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 the data will be presumably 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 5: Data management form template

Dataset identifier	The ID allocated using the naming convention outlined in Section 3.1
Title of dataset	The title of the dataset which should be easily searchable and findable
Partner	Lead partners 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.
Expected 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 will witness revisions by adding an extra set of collected data is presented in the following sub-sections.

8.2.1 VerbMobil-1Corpus dataset

Table 6: VerbMobil-1Corpus dataset

Dataset identifier	COMPRISE_Data1_VerbMobil-1Corpus_V1.0
Title of dataset	VerbMobil-1 Corpus
Partner	USAAR
Work package and task(s)	WP2: T2.1, T2.2, T2.3 WP3: T3.1, T3.2, T3.3 WP4: T4.1
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.
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)
Expected data size	9 GB
Data capture and processing methods	Unknown
Data repository	Each partner using this dataset will install a local copy once it has paid for the license. Newly collected annotations will be stored in MyBox & Zenodo
DOI (if known)	Unknown
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.

8.2.2 Integral Let's Go! dataset

Table 7: Integral Let's Go! dataset

Dataset identifier	COMPRISE_Data2_ Integral_Lets_Go! _V1.0
Title of dataset	Integral Let's Go!
Partner	USAAR
Work package and task(s)	WP2: T2.1, T2.2, T2.3 WP3: T3.1, T3.2, T3.3 WP4: T4.1
Origin	https://dialrc.github.io/LetsGoDataset/
Data description	This is a corpus of recordings of conversations on the bus information scheduling for the Allegheny County Port Authority Transit bus system via a telephone-based interface to access bus schedules and route information. This corpus is to be complemented by newly collected privacy-related annotations.
Purpose and relation to COMPRISE objectives	It is relevant to COMPRISE because personal information of the callers such as age, gender, etc. might be inferable from the data and provides the basis for privacy-related research.
Type of data	Audio recordings of conversations & 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	Raw samples (audio), .csv (transcripts)
Expected data size	715 GB
Data capture and processing methods	Unknown
Data repository	Newly collected annotations will be stored in MyBox & Zenodo
DOI (if known)	Unknown
Access	Publicly available.
Restriction on sharing	Copyleft license, copyright notice may not be removed, modifications must be clearly marked, original authors' names must not be removed.
Supporting tools	n.a
Copyright and IP management	Copyright (c) 2010-2011 Carnegie Mellon University
Licensing	Let's Go Database license.
Quality assurance	n.a.
Date of Repository Submission	n.a.
Keywords	Dialogue, telephone, bus information

Version Number	1.0
Link to metadata file	n.a.

8.2.3 RecipeBox dataset

Table 8: RecipeBox dataset

Dataset identifier	COMPRISE_Data3_RecipeBox_V1.0
Title of dataset	RecipeBox
Partner	ASCO
Work package and task(s)	WP6: T6.2
Origin	https://github.com/rtlee9/recipe-box
Data description	Structured recipes scraped from food websites. Translated versions of these recipes and/or extended instructions for the users may be added.
Purpose and relation to COMPRISE objectives	It is relevant to COMPRISE as one of the consumer demonstrator applications (cooking app giving advice during food preparation) needs a recipe database as a foundation
Type of data	Textual data
Utility	A foundation of recipes will be built which COMPRISE will use to read out to the user and instruct him/her during the cooking process
Expected reuse	The app, and therefore the used data will be used after the project lifetime as well as for individual exploitation regarding B2C business.
Type format	JSON
Expected data size	196 MB
Data capture and processing methods	Manual summarisation of recipes based on various cooking websites.
Data repository	Newly generated data will be stored in MyBox & Zenodo
DOI (if known)	Unknown
Access	Publicly available.
Restriction on sharing	No restriction.
Supporting tools	n.a
Copyright and IP management	Copyright © Ryan Lee, 2019, eightportions.com
Licensing	ODC Attribution License (ODC-By)
Quality assurance	n.a.
Date of Repository Submission	n.a.
Keywords	Recipe, cooking, instructions
Version Number	1.0
Link to metadata file	n.a.

8.2.4 Drive-Thru beta testers dataset

Table 9: Drive-Thru beta testers dataset

Dataset identifier	COMPRISE_Data4_ Drive-Thru beta testers_V1.0
Title of dataset	Drive-Thru beta testers dataset
Partner	NETF

Work package and task(s)	WP6: T6.2
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 will be used to build the Drive-thru voice-based application as described in the objectives of the COMPRISE project. For instance, a Drive-thru customer will be 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
Expected 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.

8.3 Appendix C. Metadata file template

Adding to the Zenodo metadata schema (discussed in Section 3.2), the uploaded datasets will be accompanied by a metadata file containing the information provided in the data management form in Appendix B.

Table 10: 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: