D1.2. Data Management Plan

Observing and Negating Matthew Effects in Responsible Research and Innovation Transition

Version 1.0
Public

This deliverable details the procedures for data management and the protection of personal data adopted by the ON-MERRIT consortium.

ON-MERRIT - Grant Agreement 824612

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no 824612.

Document Description

D1.2 - Data Management Plan

D1.2 Data Management Plan			
WP1 - Project Management			
Due date	31.03.2020	Actual delivery date:	27.03.2020
Nature of document	Report	Version	1.0
Dissemination level	Public		
Lead Partner for deliverable	Know-Center GmbH Graz		
Authors	Hannah Metzler, Nancy Potinka, Angela Fessl, Bikash Gyawali, Bernhard Wieser, Birgit Schmidt, Thomas Klebel, Tony Ross-Hellauer		
Reviewers	Nancy Potinka, Angela Fessl, Bikash Gyawali, Bernhard Wieser, Birgit Schmidt, Thomas Klebel, Tony Ross-Hellauer		

Revision History

Issue	Item	Comments	Author/Reviewer
V0.1	Draft version	Created based on H2020 template questions and DMP online tool	Hannah Metzler
V0.2	Revised	comments regarding T3.1	Nancy Pontika
V0.3	Revised	comments and updates regarding KC DB	Angela Fessl
V0.4	Revised	comments regarding task T3.1	Bikash Gyawali
V0.5	Revised	comments regarding WP5	Bernhard Wieser
V0.6	Revised	general feedback and comments	Birgit Schmidt
V0.7	Revised	general feedback, comments regarding task 5.2	Antonia Maria Correia
V0.8	Final internal draft	Integrating feedback and comments of all partners	Hannah Metzler
v0.9	Post-review draft	Integrated feedback from OU review	Thomas Klebel
v1.0	Submitted version	final copy-editing & formatting	Thomas Klebel, Tony Ross-Hellauer

Table of Contents

Tables

Abbreviations

DDB: Data Driven Business Team
DMP: Data Management Plan
KC: Know-Center GmbH Graz
MAG: Microsoft Academic Graph
OU: Open University
TUG: Technical University Graz
UGOE: Georg-August-Universität Göttingen
UMINHO: University of Minho

Executive summary

This Data Management Plan (DMP) contains a description of all datasets to be collected, generated or re-used as part of ON-MERRIT's research activities.

This includes, firstly, the purpose of data collection (i.e. research questions), data origin, type, sharing and utility, as well as expected file size and format. Datasets for ON-MERRIT include primary data collected via surveys, interviews and workshops, as well as compiled, processed data generated from the knowledge graph Microsoft Academic Graph, the CORE aggregator, and Google Patents Public Data.

Second, this DMP outlines measures taken to ensure compliance with the FAIR data principles.[1] Most datasets generated during ON-MERRIT will be made findable and openly accessible (whenever possible, depending on data protection requirements) via the repository Zenodo together with detailed metadata and a digital object identifier. To make data interoperable, the consortium will use Open Source software and file formats wherever possible. To enable and stimulate re-use of data, analysis code will be shared via GitHub, appropriate open licences (CC-BY or similar) will be assigned, and quality is ensured via internal review processes.

Third, the DMP describes the allocation of resources for data management, storage solutions and back-up plans during the project, long-term retention and data security, ensured via encryption and password protection. It finishes with a description of technical and organizational measures put in place for the protection of personal data, including collection and internal sharing of personal data (where necessary), anonymisation and pseudonymisation techniques used before making data openly accessible, and informed consent procedures.

This DMP is a living document that will be regularly updated by all partners. At a minimum, it will be updated once after the end of the first reporting period (30 September 2020, report deadline 30 November 2020), at the end of phase 2 (30 September 2021) and, if additional changes are required, one final time before the final review (3 March 2022).

[1] https://www.go-fair.org/fair-principles/

1. Introduction

This Data Management Plan (DMP) covers all aspects of the collection, storage, protection, and publication of datasets collected, generated or re-used within ON-MERRIT. It outlines measures taken to ensure compliance with the FAIR data principles[2] and describes the allocation of resources for data management, storage solutions and back-up plans during the project, as well as measures taken for the protection of personal data.

Chapter 2 lists the project's research questions along with the datasets which will be used to answer them. Chapter 3 details all measures taken to ensure ON-MERRIT's datasets are in compliance with the FAIR data principles. Chapter 4 summarises the resources allocated for data management. Chapter 5 lays out in detail which measures are implemented to ensure secure storage of all data used within ON-MERRIT. Finally, chapter 6 includes details about the protection of personal data. This section provides the regulations for the ethical deliverable D7.2 POPD - Requirement No. 2 Protection Of Personal Data, given that they are closely intertwined with the DMP's other aspects.

2. Data summary

The ON-MERRIT consortium will collect data to answer research questions relating to cumulative advantages (Matthew-effects) resulting from the implementation of Open Science (OS) and Responsible Research and Innovation (RRI) policies within academia, industry and policy-making. All collected data will be directly relevant to the research questions, and include no sensitive information. The various datasets generated will be used to answer questions relating to:

- The inclusion and effect of RRI and OS criteria in institutional policies.
- The effect of RRI and OS policies on researchers' access to literature and career progression depending on their institutional standing, gender and geographical location.
- The participation of researchers in RRI and OS training.
- The uptake of OS outputs (including publications, data and code) by economic actors in SMEs and large companies in Europe, in the European patent literature and in policy-making, and drivers and barriers to the uptake in these domains.
- Information seeking behaviour of economic actors and policy-makers.
- The participation of public actors and citizen scientists in evidence-gathering activities and the uptake of such research in policy-making.
- The effect of traditional and potentially new RRI and Open Science indicators on research practices.

All collected datasets will include variables describing characteristics such as gender, institutional role or standing, geographic location (at the country scale), financial resources, skills and knowledge to investigate Matthew effects along these dimensions. Data collection will focus on individuals and organisations in the case study areas of agriculture, climate and health, but will also contain data from other areas.

[2] https://www.go-fair.org/fair-principles/

Most datasets generated as part of ON-MERRIT will be made available via the certified repository Zenodo[3] on a dedicated community page[4] after appropriate anonymisation of personal information. All code generated for data-analysis will be made available via the Github repository of ON-MERRIT.[5] Due to privacy concerns, qualitative research data (interview and workshop recordings and transcripts) will not be made openly accessible. All participants will be asked for their consent before recording interviews or workshop discussions. On request we will refrain from audio recording to maximize confidentiality and openness.

Most of the data will be primary data, collected specifically for the purposes of ON-MERRIT using surveys, qualitative interviews, workshops and desk research. Some of the data will be generated from information available in public databases using computational analyses. Table 1 gives an overview of the datasets the project will generate or collect, including their type, origin, file format, and expected file size. For datasets with an asterisk in the column "Data origin", further specifications are provided below. Whenever possible, data files will be converted to file formats readable with open source software.

DS #	Task	Lead partner	Dataset (DS)	Data type	Data origin	File format	Expected size	Sharing
DS 1	3.1.	OU	Career promotion policies	Obser-vational; raw data	Human coding of publicly available policies*	.csv	<100 MB	yes
DS 2	3.1.	OU	Career profiles and research papers dataset	Derived/ compiled; processed data	Publicly available*	.csv	<20 GB	yes
DS 3	3.3	UMINHO	Survey on RRI and OS training	Obser-vational, raw data	Primary data & re-use of data from training initiatives*	.csv	<100 MB	after anonymisation
DS 4	4.1	KC, ORRG	List with bibliographic data for the literature review	Compiled	Publicly available	.docx, pdf	<10MB	yes
DS 5	4.2	KC, DDB	Survey with information seekers in industry	Obser-vational; raw data	Primary	.csv	<100 MB	after anonymisation
DS 6	4.2	KC, DDB	20-30 interview recordings with selected	Obser-vational; raw data	Primary	.mp4 .wav	max. 30* 200MB	No

[3] https://zenodo.org
[4] https://zenodo.org/communities/on-merrit
[5] https://github.com/on-merrit/ON-MERRIT

			participants from DS 5 (year 1)					
DS 7	4.2	KC, DBB	Interview transcripts from DS 6	Obser-vational; processed data	Primary	.docx .txt	<100 MB	No
DS 8	4.3	UGOE	Quantitative data on OS practices in European patent literature	Derived/ compiled; processed data	Publicly available*	.csv	20 GB	yes
DS 9	5.1	TUG	List with bibliographic data for the literature review	Compiled	Publicly available	.docx, pdf	<10MB	yes
DS 10	5.2	UMINHO	Survey with contact points in parliaments and ministries	Obser-vational; raw data	Primary	.csv	<100 MB	after anonym isation
DS 11	5.2	UMINHO	20-30 interview recordings with selected participants from DS 9	Obser-vational; raw data	Primary	.flac .mp3	max 30* 200MB	No
DS 12	5.2	UMINHO	Interview transcripts from DS 10	Obser-vational; processed data	Primary	.txt or .docx	<100 MB	No
DS 13	5.3	TUG	Recordings of 3 expert workshops	Obser-vational; raw data	Primary; not publicly available	.flac or mp3	max 900MB *3	No
DS 14	5.3	TUG	Workshop transcripts of DS 13	Workshop transcripts	Primary	.txt	<100 MB	No
DS 15	5.3	TUG	ca 20 Recordings of individual interviews with experts	Obser-vational; raw data	Primary	.flac or mp3	max 20*200MB	No
DS 16	5.3	TUG	Interview transcripts from DS 15	Obser-vational; processed data	Primary; not publicly available	.txt	<100 MB	No
DS 17	6.1	OU	Survey of academics at institutions of DS1	Observational; raw data	Primary	.csv	<100 MB	after anonym isation

Table 1. Overview of datasets.

Further specifications regarding specific datasets from Table 1:

- DS 1: PDFs of institutional promotion policies were manually collected from institutional websites. If the website requires logging into an intranet, contact persons mentioned on the websites will be asked to provide the documents via email. They were assured that no information linking their institution to the content of the policy will be published.
- DS 2: For task T3.1, we will extract data relevant to our research questions from two databases containing information about scientific publications: 1) Microsoft Academic Graph (MAG)[6] and 2) the global aggregator CORE[7]. The data extracted from these databases will be limited to the universities under study only; i.e. only a small subset of those datasets will be relevant to us in practice. The extracted data will be used to analyse the universities in terms of their publication counts, status (Open Access or not) and similar other characteristics.
- DS 3: Task T3.3. will (potentially) re-use datasets collected by training initiatives of the projects FosterPlus[8], FIT4RRI[9] and the Open Science & RRI trainer bootcamp[10], as well as data from the Open Science overview in Europe by OpenAIRE's National Open Access Desks[11] and from the Eurodoc surveys[12]. It will contact individuals involved or participating in these initiatives with a survey designed specifically for ON-MERRIT.
- DS 8: Task 4.3 will use only publicly available datasets. The main data source will be Google Patents Public Data which is openly available via Google's BigQuery service. Preliminary analysis shows that the structured full-text corpus comprises more than 800,000 patents that have been filed since 2011 from nine EU-based patent offices. However, the data is constantly updated.

2.1. Data utility

The data collected during the project may in the future be useful for researchers interested in the topic of RRI and open science within the social and computational sciences, and within Science and Technology Studies. It may further be relevant for librarians, open-science policy makers, administrative staff at research institutions, science funding organisations, as well as actors in industry interested in Open Data sources and information seeking practices.

3. FAIR data

The following sections describe which steps will be taken to make the data findable, (openly, where possible) accessible, interoperable and re-usable.

[6] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[7] https://core.ac.uk/
[8] https://www.fosteropenscience.eu/
[9] https://fit4rri.eu
[10] https://www.fosteropenscience.eu/node/2718
[11] https://www.openaire.eu/contact-noads
[12] http://eurodoc.net/oldwebsite/index-107.htm

3.1. Making data findable, including provisions for metadata

Raw data and all processed versions of data will be saved in separate folders. Dataset names will include version numbers, starting with v01 for the raw data. Datasets will be named according to the following convention, using the project name, the work package number, title, version number, year and month: ON-MERRIT_<WPNo.>_<datasetTitle>_<versionNo.>_<YYYY_MM>. An example title would be: ON-MERRIT_WP3_PromotionPolicies_v01_2020_04.

Zenodo associates a digital object identifier to each data file, making the data findable and identifiable. Metadata will be provided by filling in the respective fields of the interface on Zenodo. Fields with an asterisk are mandatory, but all fields listed below will be completed for ON-MERRIT's datasets.

- Publication date*
- Title*
- Authors* (include ORCID for all authors)
- Description*: This section will be used to describe the origin and nature of the data, and details concerning its potential future users. It will provide the link to the Github repository hosting the analysis code. All metadata necessary to interpret and re-use the data will be provided, including:
    - the number of variables and data points (participants, papers, policies etc.)
    - names and explanations of variables
    - data format
    - version of used software
- Version: the version number of the dataset, if several versions are uploaded
- Language: English
- Keywords: The keywords included in the publication/report presenting the dataset will also be provided as search keywords for the dataset, in order to optimize possibilities for re-use of the data.
- Additional notes:
    - If not included in an openly available report/publication (i.e. Open Access), details about procedures of data collection and analysis should be provided in this section.
    - Also provide the following funding information here, if you experience difficulty inserting it in the Funding tab: "The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no 824612."
- Access rights*: This will in general be "Open Access" for ON-MERRIT, if data has been anonymised and the report published or the article publication submitted for publication.
- License*: we will use at least the Creative Commons Attribution 4.0 International licence (or equivalent), and prefer Public Domain Label (e.g. CCO[13]) with a recommendation to cite data creators.
- Funding: Insert the European Commission (EU) and the Grant Agreement number 824612 in a dedicated "funding information" metadata field if possible, otherwise insert it in the Additional notes tab.
- References: Provide a reference to the report and/or scientific publication that used the data and describes the collection and analysis procedures in more detail. A persistent identifier to such publications must be included if available.

[13] https://creativecommons.org/share-your-work/public-domain/cc0/

3.2. Making data openly accessible

After appropriate anonymisation (see section 5.3) according to the General Data Protection Regulation (GDPR)[14], most datasets will be made available via Zenodo, which offers 50GB of storage space per dataset and imposes no size limit for community accounts. Whenever possible, open datasets will be shared in a format accessible with open source software. Used software includes (1) for quantitative data analysis: Python, R, Git, SPSS and MS-Excel, (2) for analysis of qualitative data: MAXQDA, (3) for text files: MS-Word, Adobe Reader, PDF-Xchange, and (4) for reference management: Zotero. All of these are commonly used for research purposes, and extensive documentation is available online. All software except SPSS, MS-Word and -Excel, MAXQDA is open source. No proprietary software is needed to re-use our data, and all used proprietary software is commonly used for research purposes.

3.3. Making data interoperable

To allow data exchange and re-use, all data will be saved in the most interoperable format possible (.txt or .pdf for text data, .csv for tabular data, .R and .py for code, .tif, .png, .svg, .jpeg for images, .flac, .wav for audio data, see the EPFL Library's Fast Guide for Research Data Management[15], p. 4). The analysis in tasks 3.1, 3.2, 4.3 and 6.1 will follow the idea of a research compendium[16], where the analysis is shared along with the code and data. A research compendium allows definition of the computational environment, increasing the computational reproducibility of the analysis. We will provide metadata using the standard format of the repository Zenodo as detailed in section 2. 1.

3.4. Increase data re-use

Data will be made available for re-use with at least a CC-BY licence at the latest upon the publication of research results. This means it will be published together with the release of preprints if possible, and at latest on publication of the peer-reviewed article analysing the respective data. Most of the data produced in the project will be fully available for others to re-use. Only data including personal information according to the GDPR will be restricted. All openly accessible datasets will remain re-usable for at least 20 years, the time period for which the repository Zenodo is currently financed.

3.4.1. Procedures for quality assurance

The ON-MERRIT consortium will ensure that all shared data is consistent with quality standards required to publish in peer-reviewed journals. To the best of our abilities, we will describe the data collection and analysis procedures in sufficient detail to allow reproducibility by trained professionals in the respective project reports or research papers. All software code developed for this project will be hosted on https://github.com/on-merrit/ON-MERRIT which follows the Cookiecutter Data Science project

[14] https://gdpr-info.eu/issues/personal-data/

[15] https://www.epfl.ch/campus/library/wp-content/uploads/2019/09/EPFL_Library_RDM_FastGuide_All.pdf

[16] https://research-compendium.science/

structure[17]. This is a standard template for data science projects which facilitates reproducible research and has been used in a number of other software projects.

As a general measure of quality assurance, all analyses and deliverables undergo an internal review process involving at least two assigned consortium members. All survey instruments will be tested before being sent out to participants, and data analyses will be performed in a team consisting of at least two professional researchers. Similarly, the coding of the interview transcripts will be conducted by at least two professional researchers using a corresponding tool like QDA Lite. We will develop a coding schema to address the research question assessed in the interview and calculate intercoder reliability, to capture in how far the different independent coders agree in their application of the coding scheme.

4. Allocation of resources

The primary responsibility for data management in ON-MERRIT lies with the project coordinator. However, each task leader is responsible for handling the data they collect as part of their tasks according to the regulations and following the procedures laid out in this DMP. One exception is the storage of survey data, which will be handled centrally by the coordinator using storage resources at TUGraz and Know-Center.

Writing and updating the data management strategy is estimated to require personnel costs corresponding to 15 working days for the project coordinator over the entire project duration of 30 months, and 3 days for partners. For the documentation of data that allows FAIR re-use, for anonymization, copyright assessment, data cleaning and publishing, as well as for preparing data for long-term preservation in a data repository, an estimated 1-10 days will be needed per publication/dataset, with bibliographic data requiring less and computational data accompanied by code that requires extensive documentation requiring more time. The required personnel costs for data management are covered in the project costs budgeted for ON-MERRIT.

Institutional servers at partner institutions require no additional cost, and are partially covered by the Overhead costs included in the grant agreement. The data repository Zenodo used for long-term preservation of the data is free of charge. No additional project related costs are necessary for usage of existing institutional infrastructures to store personal data.

5. Data security

During the project, a copy of all data acquired by consortium members at Know-Center and TU Graz will be stored via Nextcloud, for which both institutions offer 50GB per employee, or other internal institutional servers including KnowNew at Know-Center. All data stored via these services is backed up daily to two or three physically separate and secure storage locations at Know-Center and TU Graz, respectively. Access to the cloud is only possible with valid credentials and access rights for the folder in

[17] https://drivendata.github.io/cookiecutter-data-science/

question, and from within the institution's network or using its VPN channel. For transfer to the cloud, data is encrypted. On the cloud, all datafiles containing personal data as defined in the GDPR (see section 5.1) will be put into an encrypted folder (protected with a password). Interview and workshop recordings (audio files) will only be stored in encrypted folders and on these internal institutional network drives. Transcripts may be shared within the consortium for internal review purposes via Nextcloud if necessary.

University of Minho uses an institutional repository with a regular backup schedule. The University of Goettingen uses a self-hosted GitLab[18] instance and the "Goettingen Research Online"[19] data repository as institutional backup solutions with a regular backup schedule. At Open University, data is stored on at least one of the following servers:

- Open Research Data Online (ORDO) Repository: Daily backups and kept for 5 days. Weekly snapshots taken of the entire data system.
- OneDrive: Files backed up by Microsoft and copies held on multiple Microsoft servers in multiple locations within the EU.
- OU network file storage: Regular backups taken according to best practice.
- SharePoint: Regular backups taken according to best practice.

During the project, consortium members may store project data on their laptops with appropriate password protection. To ensure regular backup of this data, it has to be stored within a secure network drive (Nextcloud, KnowNew etc).

After the end of the project, the use of Zenodo as a certified repository[20] for data, reports and articles will ensure long term preservation and curation of data (20+ years)[21]. Personal data which are not made openly accessible will be stored locally on the institutional servers of each partner institution, and thus be available for at least 10 years at Know-Center and TU Graz, and 5 years at Open University. Interview and workshop recordings are an exception to this and need to be destroyed after publication of the results (i.e. the reporting of results directly related to the research in ON-MERRIT). Transcripts may be stored for the full retention period (10 or 5 years, depending on the institution). At Know-Center, data deletion after the retention period is ensured by placing the data in the folder "knownew/900 ARCHIV/after year - 10" on the KnowNew Server at the end of the project. University of Minho will store processed data on the repository Dataverse for long-term preservation, which supports restrictions to limit the use of or access to data that cannot be publicly accessible and provides a backup copy for safekeeping.

5.1. Encryption and password creation

For encrypting data files or folders containing personal information, we will use the 7-ZIP software, for which a step-by-step guide is described below.

[18] https://www.gwdg.de/e-mail-collaboration/gitlab
[19] https://data.goettingen-research-online.de/
[20] https://www.coretrustseal.org/why-certification/certified-repositories/
[21] https://about.zenodo.org/policies/

5.1.1. Step-by-step guide for encryption

This guide explains archiving and password protection with the software 7.zip:

- Go to 7-ZIP website[22] and download an appropriate installation file
  - for Windows users it is recommended to download the latest version of x64-bit software (.exe or .msi file)
  - for Linux user; find the 7-Zip in Software store (GUI version) or install it with the CLI command: sudo apt install p7zip -y
- After the installation of the software, select the file/folder you wish to archive (and password protect), right click on it and select:
- 7-Zip → Add to Archive
- new window will pop up
- in Archive section choose the name and the location were the ZIP file will be saved
- everything else should be left as default
- in Encryption section choose your password: this is the password you will use to decrypt and unpack the archive later on, so save it in your password manager or remember it. Consider following the password created guidelines below.
- select OK and archiving/encrypting process will start
- If no password is provided, the archived file will not be password protected!
- To access the files in an encrypted folder: Right click – 7-Zip – Open archive. Enter password upon double clicking the file you want to open.
- In order to ensure that the used password is secured, store it in a password manager program like the open source software Keepass (https://keepass.info/)

5.1.2. Password creation

Referring to the strong password recommendations of the Ubuntu Community, all passwords should be reasonably complex, unique to the employee and difficult for unauthorized people to guess. Therefore, the following general rules apply:

- Avoid (complete) dictionary words, common phrases and names.
- Where possible, do not use the same password for various access needs.
- Do not use a password for company accounts that you already use for a personal account
- Do not base your password on personal information (e.g. names of family members, birth date, names of pets, license plates, phone numbers, username etc) or other well-known or easily accessible information.
- Do not use word or number patterns like aaabbb, zyxwyuts, 123321, 1234abcd etc.
- Do not use any of the above preceded or followed by a digit (e.g. secret1, 1secret etc.)
- Use a password that is significantly different from your previous password.

[22] https://www.7-zip.org/download.html

6. Protection of Personal Data

Regulations regarding identification and recruitment of research participants for the data collection during ON-MERRIT are detailed in the ethical deliverable D7.1 H - Requirement No. 1 Human participants. Regulations for the ethical deliverable D7.2 POPD - Requirement No. 2 Protection Of Personal Data are provided in this section of the DMP, given that they are closely intertwined with other aspects of this deliverable. In the ethics self-assessment during the application process, the Project Coordinator mistakenly answered the question about whether ON-MERRIT involves further processing of previously collected personal data (secondary use) with yes (see the Ethics Summary Report). Given that the proposal itself mentioned no secondary use of personal data, the Post-Grant Requirements included a request for clarification regarding the secondary use of data in the process of doing Scientometrics, Text Mining and Social Network Analysis. Actually, the ON-MERRIT consortium will only use these methods to analyse publicly available data (e.g. patent data, data extracted from MAG, CORE) that contains no personal information. Nevertheless, the following section details measures taken to protect personal data that is collected for the purposes of the project (primary use).

6.1. Definition of personal data

The ON-MERRIT consortium refers to the GDPR definition of personal data as follows: "Personal data is any information that relates to an identified or identifiable living individual. Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data. Personal data that has been de-identified, encrypted or pseudonymised but can be used to re-identify a person remains personal data and falls within the scope of the GDPR. Personal data that has been rendered anonymous in such a way that the individual is not or no longer identifiable is no longer considered personal data. For data to be truly anonymised, the anonymisation must be irreversible."[23] Examples of personal data are a name and surname; a home address; an email address such as name.surname@company.com; an identification card number; location data; an Internet Protocol (IP) address; a cookie ID; the advertising identifier of a phone. In contrast, data such as a company registration number, and anonymised data are not considered personal data. Additionally, the GDPR does also not apply to information "about legal entities such as corporations, foundations and institutions. […] Data must therefore be assignable to identified or identifiable living persons to be considered personal."[24]

6.2. Collection and sharing of personal data in ON-MERRIT

Personal data collection will be kept minimal, involving only data directly relevant to the research questions of ON-MERRIT. To the extent that the research methodology and research question allow, data will be collected in an anonymous form from the beginning. During interviews, workshops and in surveys, email addresses will only be collected on a voluntary basis, i.e. if participants wish to be informed about research results. These email addresses will be deleted at the end of the project, and never shared publicly. Apart from the email address, ON-MERRIT does not require the collection of any other direct

[23] https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en

[24] https://gdpr-info.eu/issues/personal-data

(e.g. name) or strong indirect identifiers (e.g. national insurance number), but only of variables that may reveal the identity of an individual in combination (see section 5.3). After data collection, such personal data will be kept confidential, i.e. access to will be restricted to authorised researchers in the consortium. Any dataset including personal data will only be shared with partners after encryption, with passwords shared across communication channels separate from the data sharing channel, such as during meeting calls or via the open source software Passbolt[25].

Data containing personal information will only be shared openly via Zenodo after ensuring that individual research participants cannot be identified (see section 5.3.). Interview and expert workshop data will not be made openly accessible, as recordings of interviews and expert workshops and their transcripts (DS 6 & 7, DS 11-16) necessarily contain detailed personal information. Due to this level of detail, full non-identifiability cannot be guaranteed even after blinding of personal information.

6.3. Anonymisation and pseudonymisation techniques

Before sharing data openly, ON-MERRIT consortium will anonymise data to make participants unidentifiable. Anonymisation implies removing both variables that directly identify the individual (in ON-MERRIT: name, email address), and variables that indirectly identify the individual (in ON-MERRIT: combinations of country, name of organisation, gender, seniority etc.)[26]. To adapt the level of anonymisation to the sample size of each dataset we will use the k-anonymity technique as implemented in the tool Amnesia[27] by OpenAIRE (online or downloadable version available, instructions here[28]). Alternative tools that may be chosen according to the preference of each partner are ARX[29] and µ-ARGUS[30]. The specific anonymisation procedures that are eventually applied (which variables are deleted, re-grouped etc) will have to be updated in later versions of this DMP.

Where full anonymisation is not possible, pseudonymisation will be used. Re-contacting survey participants for follow-up interviews requires storing their email address until completion of the interview. For this purpose, the ON-MERRIT consortium will use pseudonymisation, that is, keep only a code to identify the participant (the subject identifier) in the survey datafile (the main data set), and keep personal data (e.g. the email address) with a copy of the subject identifier in a separate encrypted file (the identification dataset). The identifier may simply consist of a subject number (S01, S02, etc) which the researcher collecting the data assigns to each participant during or after data collection. As soon as the subject identifier has been assigned, all single-column variables identifying the participant have to be deleted from the main dataset, and only be kept in the encrypted identification dataset.

After this pseudonymisation procedure, personal data will only remain in the form of combinations of variables in survey data that could be used for direct or indirect identification of individuals, e.g. in the surveys with SMEs, parliament and ministry staff, researchers at specific institutions. In these cases, data

[25] https://www.passbolt.com

[26] https://www.fsd.tuni.fi/aineistonhallinta/en/anonymisation-and-identifiers.html#terms-to-understand

[27] https://amnesia.openaire.eu/amnesiaInfo.html

[28] https://amnesia.openaire.eu/documentation.html#fh5co-tab-feature-vertical1

[29] https://arx.deidentifier.org/downloads/

[30] https://arx.deidentifier.org/downloads/

will be stored in encrypted folders on internal institutional servers (see section 4). Before sharing such data openly, the necessary variables will be cut until the risk of identifiability is sufficiently low before making data openly accessible. This also requires a consideration of the sample size, as individuals are more easily identifiable in small samples. Such combinations of variables may specifically include gender, name/type of the organisation/institution, role or seniority of the participant, and geographic location which will prospectively be collected in the following datasets:

- The authors and research papers dataset (DS 2) includes name and affiliations of authors as well as identifiers used for them in MAG and/or CORE datasets. This information will be used for making predictive analysis but they will be removed until anonymity is ensured before making the data openly accessible.
- DS 5: The survey with information seekers in SMEs conducted in T4.2.
- DS 10: The survey with policy-makers in parliaments and ministries conducted in T5.2 as well as follow up interviews will collect person's names, affiliations, countries and email addresses; this dataset cannot be public, but will be described and stored in Dataverse.
- DS 17: The survey with researchers at specific institutions conducted in T6.1.

6.4. Informed consent procedures

Only adult research participants above 18 years old will be eligible for participating in ON-MERRIT's research activities. Written informed consent for processing, sharing and long-term preservation of anonymized data will be collected at the start of online surveys, face-to-face interviews, and expert workshops. Research participants will be informed about the type of data collected (audio, transcripts, online survey answers). The consent form will be provided together with an information sheet that explains/mentions:

- the subject and objectives of the research,
- data collection methods,
- the voluntary nature of participation,
- confidentiality,
- the possibility to withdraw from the project at any moment without any justification,
- contact information of the researcher responsible for the study
- information about the potential reuse of data: sharing of anonymised/aggregated data on data repositories for re-use by other researchers

A template for the participation information and a consent certificate specific for each type of research activity will be provided in the annex of the ethics deliverable D7.1 H - Requirement No. 1 Human participants. Each partner is responsible for storing physical versions of consent forms collected as part of their tasks for at least two years after the end of the project, the time period during which the European Commission may order an audit. Electronic versions should be kept along with the data until its deletion in an encrypted folder (see section 4).

7. Other institutional regulations

All partner institutions require their employees to enter publications to an internal publication database or institutional repository. In addition, the partners listed below have to follow other institutional procedures, which are in accordance with the present Data Management Plan.

- University of Goettingen's guidelines for research data management require a DMP for each research project, and data storage in institutional or disciplinary repositories, (Forschungsdaten-Leitlinie der Universität Göttingen (einschl. UMG) 2016[31]), criteria which are covered by regulations of the present DMP.
- University of Minho: Data must be uploaded to the data repository Dataverse.
- Open University: Data could be deposited in the institutional data repository. The Open University's Open Access policy is general and complies with European Union's Open Access policy.

[31] https://www.uni-goettingen.de/de/01-juli-2014-forschungsdaten-leitlinie-der-universitaet-goettingen-einschl-umg/488918.html