# Enhancing Trust, Integrity, and Efficiency in Research through Next-Level Reproducibility Impact Pathways

## Deliverable D1.2 – Data Management Plan

**30 June 2023**

Lead Beneficiary: ARC

Author: Elli Papadopoulou
Reviewers: Hajira Jabeen, Thomas Klebel

Funded by
the European Union

**Prepared under contract from the European Commission**
Grant agreement No. 101094817

EU Horizon Europe Research and Innovation action

| | |
|---|---|
| Project acronym: | **TIER2** |
| Project full title: | **Enhancing Trust, Integrity, and Efficiency in Research through Next-Level Reproducibility Impact Pathways** |
| Start of the project: | January 2023 |
| Duration: | 36 months |
| Project coordinator: | Dr. Tony Ross-Hellauer |
| | |
| Deliverable title: | Data Management Plan |
| Deliverable n°: | D1.2 |
| Version n°: | 1.1 |
| Nature of the deliverable: | Report |
| Dissemination level: | Public |
| | |
| WP responsible: | WP1 |
| Lead beneficiary: | ARC |

TIER2 Project, Grant agreement No. 101094817

Due date of deliverable: Month n°6
Actual submission date: 30 June 2023

Deliverable status:

| Version | Status | Date | Author(s) |
|---|---|---|---|
| 0.1 | Draft | 05 June 2023 | Elli Papadopoulou<br>ARC |
| 0.6 | Draft | 21 June 2023 | Input from TIER2 Consortium |
| 0.7 | Review | 21 June 2023 | Hajira Jabeen (GESIS)<br>Thomas Klebel (KNOW) |
| 1.0 | Final Version | 28 June 2023 | Elli Papadopoulou<br>ARC |
| 1.1 | Final review | 30 June 2023 | Thomas Klebel & Tony Ross-Hellauer<br>KNOW |

# Table of contents

# Executive Summary

This deliverable reflects the data management activities of the TIER2 project. TIER2 will develop next-level reproducibility tools, practices & policies across diverse epistemic contexts to increase trust, integrity, & efficiency in research. In this context, TIER2 will itself adhere to radical reproducibility & transparency to ensure best practices, including adherence to Horizon Europe requirements on Research Data Management & Open Science. At the meta-level, the DMP of the project will progressively incorporate elements of reproducible research to realise a prototype of a new concept towards "Reproducibility Management Plans (RMPs)". This enhancement will be equally supported by the development of the Reproducibility Management Plan tool that is expected to be completed over the course of the project's lifetime.

The TIER2 DMP will be treated as a "living document" that will be continuously updated to record progress and changes in the decisions of the data management and reproducibility practices followed by the consortium. This first version of the DMP as well as its future iterations in M18 and M36 will be linked and available as machine actionable and FAIR outputs produced by ARGOS service (argos.openaire.eu): 10.5281/zenodo.8092430.

The structure of D1.2 "Data Management Plan" follows the European Commission's Horizon Europe Data Management Plan Template topics and answers the contained questions with information that members have in this initial phase of the project. It should be noted that answers relevant to Section "3. Other Outputs" from the EC's template are embedded in all sections where relevant, regarding software and code. At the end, we provide specific examples of data that the project is / will be generating, collecting or reusing in the form of tables.

# List of Abbreviations

EU – European Union
DMP – Data Management Plan
EOSC – European Open Science Cloud
EC – European Commission
DoA – Description of Action
WP – Work Package
European Research Area – ERA
Open Science Framework – OSF
maDMP – machine actionable DMP
CV – Controlled Vocabulary

# 1. Data Summary

The DoA has been the initial point of reference for the first version of the TIER2 DMP. It describes all activities to be performed in the TIER2 lifetime dividing the work into packages of concrete goals and objectives, expected outcomes and outputs and responsible consortium partners. The data, software and other research output management activities and any dependencies in the communication and coordination of efforts between partners were identified by examining the DoA. These are presented below:

- WP1 Coordination and Management
    - Task 1.2 Financial Coordination: should be informed about the data management practices followed by the consortium to ensure eligibility and better allocation of costs on data management.
    - Task 1.3 TIER2 Open and reproducible research practices: will produce a dataset by undertaking an autoethnography study to enhance the DMP with reproducibility practices.
- WP2 Communities, Communication and Dissemination
    - Task 2.2 Community development and coordination of co-creation activities: will produce datasets from the open call to build the network of Reproducibility Networks and from virtual brainstorming events or “BarCamps” to co-create whitepapers on topics such as needs-gap analyses, barriers & enabler assessments, & virtual “co-working” events or “hackathons” to promote & improve reproducibility tools developed & piloted in WPs4/5.
    - Task 2.3 Development of the Reproducibility Hub: will develop a platform that will take as an input existing datasets while classifying in its content other types of useful resources, such as training material for reproducibility.
- WP3 Concept, Evidence, Synthesis and Recommendations
    - Task 3.2 Evidence-base and inventory of reproducibility tools and practices: datasets will be reused and desk research will lead to derived datasets. The outputs of this activity will later become inputs of the Reproducibility Hub (T2.3).
    - Task 3.3 Synthesis and recommendations: will produce datasets by synthesizing results from the pilots & survey co-creation communities to support recommendations according to the Delphi methodology.
- WP4 Community-Driven Design and Piloting of Reproducibility Tools and Practices
    - Task 4.1 Future studies to identify priorities from the stakeholder community to predict future of reproducibility and identify actionable steps: will collect audio and generate a transcribed dataset from getting input from participants during the online scenario workshops.
    - Task 4.3 Pilots preparation activities & Task 4.4 Pilot implementation and assessment: will produce datasets corresponding to the pilot activities.
- WP5 Development of Tools and Practices for Communities: might, progressively, involve software management apart from data management activities.

We further detail the specific datasets and activities by answering the Horizon Europe DMP template questions in the following sections of the deliverable.

## 1.1. Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

TIER2 uses OpenAIRE, FAIRsharing and GESIS as data providers to further exploit their content from the perspective of reproducible science. Specifically, data included in the OpenAIRE Graph and FAIRsharing will be (re)used, enhanced and contextualised to support the development and content enhancement of the Reproducibility Hub (T2.3). The Hub will be available in the form of a wiki-based web-based platform and serve as the knowledge base of reproducibility practices and tools. Similarly, GESIS datasets will be selected to support activities linked to the pilots that the project will perform. The pilots will specify new interventions to increase reproducibility across all phases of the research lifecycle from ideation to assessment for different methodologies and epistemic contexts and they will support the enhancement of existing and the development of new tools and services.

## 1.2. What types and formats of data will the project generate or re-use?

Most of the activities that will be performed in TIER2 leading to the project deliverables and results stem from scoping, enhancing and assessing reproducibility in research for different stakeholders, domains and at different levels, from theoretical to practical applications. The dataset types and formats are an extension of those activities characterising their nature and scientific domains. The following are some examples:

- The types of collected data from landscaping and co-creation activities (desk research mappings, empirical studies, interviews, etc) are expected to be in tabular and text formats, e.g. comma separated values and word documents. For transcribed content, abiword formatted files will be curated and converted to more open solutions, such as OpenDocument format.
- The reused data from OpenAIRE Graph, FAIRsharing and GESIS are provided in different formats to be exploited also programmatically, such as .json .xml, .csv and .tar.

*See also Section 6 (Dataset Descriptions) to view the types and formats of data per individual dataset.*

## 1.3. What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The main purpose of collected, generated and reused data is to support the project's scoping (O1), co-creation (O2 & O5), piloting (O3) and assessment (O4) activities that will drive change and equip stakeholders with enhanced skills and tools in reproducible research (O6).

Collected data will produce feedback and validation of TIER2 project findings supporting the recommendations for science policy-makers, and boost project deliverables and publications, such as:

- a preprint of the conceptual framework for reproducibility across contexts,
- the pre-registration of the protocol for future studies,

- the pre-registration of the methods for pilot implementation/assessment
- the project self-assessment report that feeds into recommendations for how to organise international, multidisciplinary projects to foster reproducibility,
- the integrative review of the literature surrounding reproducibility of qualitative methods.

At the same time, reused data will power interactive graphs that visualise the landscape of reporting standards & best practices (for data, metadata & software), & their relations, as well as their use (by the EOSC clusters) & their adoption by data policies (by funders & publishers).

## 1.4. What is the expected size of the data that you intend to generate or re-use?

The aggregated size of managed data in TIER2, so far, is estimated to be more than 300 GB. Out of all TIER2 data, the largest in size are the reused datasets, occupying more than 250GB, as they are derived / compiled data from many data providers. The data collected or generated are smaller in size, consisting of files that sometimes do not exceed 10MB.

*See also Section 6 (Dataset Descriptions) to view the exact size per individual dataset.*

## 1.5. What is the origin/provenance of the data, either generated or re-used?

The data that are managed in TIER2 form a mixture of primary data, i.e. directly assembled data or information for the first time, and secondary data, i.e. has already been collected through primary sources and made readily available to other research(ers).

Project generated data derive from a collection of information by means of desk research, quantitative and qualitative methods that are tied to the creation of the framework, the provision of recommendations, and the co-creation activities performed in TIER2.

Reused data come from GESIS and the thousands trusted sources that are harvested, curated and contextualised in the OpenAIRE Graph and FAIRsharing. Examples include institutional and national literature and data repositories, journal databases, registries (e.g. ROR, ORCID), funder databases, other content aggregators (e.g. WoS, Scopus, OpenAlex, Crossref, Datacite), etc. For the complete list, please visit https://graph.openaire.eu/docs/ and https://fairsharing.org/search?fairsharingRegistry=Database.

*See also Section 6 (Dataset Descriptions) to view the origin/provenance per individual dataset.*

## 1.6. To whom might your data be useful ('data utility'), outside your project?

As data are incremental components for research integrity and reproducibility, their availability is important to everyone working in the field of research because they provide evidence about and validate the project findings and outcomes in a transparent and participatory fashion. Yet, as shown in the table below, TIER2 data are of immediate use by its stakeholders, including social, life, computer science researchers, publishers & funders.

**Table 1:** *Utility per stakeholder group.*

| Stakeholder | Utility |
|---|---|
| Research Funders (RFOs) | Merge with other data to enhance collected data and adapt findings related to reproducibility adoption by funders outside the consortium; support evidence policymaking |
| Publishers | Compare and merge with own data; enrich current practices and tools in support of reproducible processes and workflows |
| Researchers | Merge and/or compare with own data to provide insights on different aspects of reproducible science in their domains; provide input to (new) tools; support own practices and design of reproducibility pathways |
| Reproducibility Networks | Expand project activities and findings based on collected data and identified gaps; communicate lessons learnt; provide input in support of (new) reproducibility activities, incl. trainings |
| General public | Get informed about reproducible science to increase citizens' participation and trust in science |

The full list of TIER2 stakeholders will be available as part of the Task 2.1 deliverable that is dedicated to stakeholder mapping.

*See also Section 6 (Dataset Descriptions) to view data utility per individual dataset.*

# 2. FAIR data

Guided by the FAIR principles [2], the TIER2 consortium will employ all the necessary mechanisms and workflows to follow best practices that enrich the European Research Area (ERA) and EOSC with scientific content that is findable, accessible, interoperable and reusable. Below, we provide our collective answers with examples on specific topics addressing the questions of the Horizon Europe DMP Template.

*For more detailed information concerning FAIR application per individual dataset, please consult Section 6 (Dataset Descriptions).*

## 2.1. Making data findable, including provisions for metadata

### 2.1.1. Will data be identified by a persistent identifier?

All TIER2 data, in their processed form, will be published and assigned persistent identifiers according to the Digital Object Identifier (DOI) system via Zenodo that serves as the main project repository. New resources that are published as project deliverables, e.g. the Reproducibility Checklist (T4.1), will mint a DOI during deposition, while reused data are already findable by their assigned DOIs.

### 2.1.2. Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

All project data and published resources will be accompanied by descriptive metadata that follow the OpenAIRE (https://guidelines.openaire.eu/en/latest/) and Dublin Core/Datacite (https://schema.datacite.org/) schemas to enable their uninterrupted exchange and search for retrieval, at minimum by: title, description, author, identifier, publisher, date.

### 2.1.3. Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Keywords will be offered from all venues of publication that TIER2 will be exposing content. Data deposits and publications of project deliverables and resources will contain free text keywords in the metadata consisting of specified terms about the content, contributors, and enablers (acknowledgments) of the given outputs. Specific attention will be given to keywords that complement general metadata and support decisions on the use and reuse of data. Additionally, communication activities that target the promotion of TIER2 outputs will highlight data, software and other research outputs on the website. In support of this, the Dissemination report forms for datasets used within the consortium will include a field dedicated for keywords.

### 2.1.4. Will metadata be offered in such a way that it can be harvested and indexed?

All types of project publications, either being project deliverables and results or formal scientific papers and the data underlying them, will be described by metadata. Zenodo will host project outputs in a dedicated community, and FAIRsharing will register any newly-developed databases

and standards to improve their discoverability. The Open Science Framework (OSF) will be used for preregistrations and deposit of corresponding project datasets.

The aforementioned platforms have mechanisms to facilitate greater ranking of data and results from search engines and their ranking on the web, especially through Zenodo's integration with Google Dataset: https://datasetsearch.research.google.com/. There is a RESTAPI that can be used to satisfy such interactions: https://developers.zenodo.org/. As regards academic networks, Zenodo and FAIRsharing are harvested by OpenAIRE and enrich its Graph with content and links/relationships for research, incl. the EOSC.

## 2.2.Making data accessible

### 2.2.1. Will the data be deposited in a trusted repository?

A Zenodo community was created to serve the self-archiving needs of the project: https://zenodo.org/communities/tier2/. Zenodo is one of the four repositories that offer a complete set of metadata that are mandatory to the Horizon Europe requirements [1]. It is a trusted repository, as defined by the EC, because it fulfils all the essential characteristics required - policy, (open) access and PID assignment, metadata requirements. FAIRsharing is also trusted as it follows best practices for metadata and their curation while being endorsed by the community of Life Sciences as part of the ELIXIR Recommended Interoperability Resources, selected by external reviewers.

In addition, when appropriate, the consortium will seek thematic repositories to deposit datasets of disciplinary interest, especially those linked to the pilot activities. In this context, GESIS will assume this role for social sciences data, while FAIRsharing and re3data registry (https://www.re3data.org/) will be utilised for the selection of a trusted repository for life sciences and computer sciences.

### 2.2.2. Have you explored appropriate arrangements with the identified repository where your data will be deposited?

TIER2 organises its data archiving activities with the support of its partners UOXF and OpenAIRE which provide their services of FAIRsharing and Zenodo respectively, to the whole consortium, even expanding to externals from co-creation activities. They are directly responsible for making the appropriate arrangements with their repositories as per the grant agreement and their commitments to the rest of the consortium. For example, the policy of Zenodo limits the upload per dataset to 50GB, which in the context of the TIER2 can be surpassed.

### 2.2.3. Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

All identified repositories use Datacite as their PID provider to mint DOIs for deposited outputs. From the DOI resolver, they are then able to resolve the identifier to a digital object. In the case of Zenodo, to create related identifiers with other outputs, the repository maintains the following list of PID resolvers: https://github.com/inveniosoftware/idutils/blob/d29102410bd26be48dcac40d688659e2d19a7572/idutils/__init__.py#L962-L987.

#### 2.2.4. Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

All data will be made openly available in their fully processed and/or analysed form. Data containing personal or sensitive information will be anonymised prior to their sharing, to ensure de-identification even if reverse engineering is enforced. For example, workshop recordings (raw data) will be transcribed (partially processed data), anonymised (further processed data), analysed and made available for (re)use.

#### 2.2.5. If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

All publishing venues selected by the consortium will be fully Open Access, with preference for venues practicing open peer review where possible, thus posing no delays in the immediate access of publications and offering greater transparency in the process. Special attention is given on immediate and, on some occasions, early data sharing. For that, the open access policies and data agreements with scientific publishers will be carefully reviewed before the decision to publish in these venues will be made. Open Research Europe (ORE) is among the lists of appropriate publishing venues, with data notes, i.e. “brief descriptions of quantitative or qualitative datasets that promote the potential reuse of research data and include details of why and how the data were created” supporting the FAIR principles.

#### 2.2.6. Will the data be accessible through a free and standardized access protocol?

Metadata of data will be accessible through Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) endpoints via Zenodo, FAIRsharing and GESIS repositories.

#### 2.2.7. If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

During the project, data and intellectual works of the consortium are stored in Know-Center Teams workspace and in partners’ institutional cloud providers, especially when sensitive data (including survey data, interview transcripts, etc.) are involved. On the occasion of sensitive data, access will be restricted with passwords.

Upon completion of activities during the project as well as after the project finishes, published outputs, incl. anonymised datasets and metadata records, will be available in open access via the repositories following their self-archiving and retention policies. If data cannot be shared openly, contact details will be provided fie externals to request access to data.

#### 2.2.8. How will the identity of the person accessing the data be ascertained?

To secure the identity of the people accessing the data, repositories provide a layer of Authentication and Authorization (AA) to their content supported by AA infrastructure providers, such as EduGain, OpenAIRE, EOSC etc. The two actions are important together as, on the one hand, authentication verifies the identity of the user or service and provides reusable credentials while, on the other, authorization determines and stores their access rights. During data processing when data will be stored in institutional cloud providers, access will be provided only to task members and with passwords.

#### 2.2.9. Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

No. Raw data containing personally identifiable data will not be shared.

#### 2.2.10. Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

As per the grant agreement, metadata will be made openly available in the public domain under CC0 license. That criterion is already satisfied by TIER2 selected repositories that are exposing their content to other providers.

#### 2.2.11. How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

More information will follow in the sustainability plan of the project, but the deposited data will remain available and findable for as long as the repositories and metadata harvesters/aggregators operate. No retraction of access is expected from the consortium partners.

#### 2.2.12. Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

As regards software that supports TIER2 reproducibility activities, either developed, extended or reused, it will be added on a dedicated GitHub page, to record a collection of open-source researcher reproducibility toolsets. It should be noted that almost all software of the tools and services of the pilots is open source.

Particularly for T4.1 that will conduct 3 cross-stakeholder focus groups & 14 interviews, the NVivo software (https://lumivero.com/products/nvivo/) will be used to perform bottom-up coding of transcriptions.

### 2.3. Making data interoperable

#### 2.3.1. What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data

#### interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

Research reporting standards listed in the FAIRsharing registry will be used in the development of new reporting guidelines.

#### 2.3.2. In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

The TIER2 consortium does not foresee the creation of a new type of Controlled Vocabulary (CV). On the contrary, domain specific and community endorsed CVs will be (re)used, such as the ones mentioned in Section 6.

#### 2.3.3. Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

All datasets that will become underlying datasets of TIER2 publications, will carry related identifiers that show relationships with the given publication(s), other datasets that they might be parts of and any software relevant to their processing and handling. This is possible on Zenodo (related_identifiers) and on the ma-DMP version of TIER2 DMP on ARGOS by utilising its semantics.

### 2.4.Increase data re-use

#### 2.4.1. How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Documentation that supports data analysis validation and reuse will be made available upon data deposit as accompanying materials. The consortium has already identified the use of readme files and codebooks among those practices.

#### 2.4.2. Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

To the extent possible, all processed and anonymised datasets will be made available under Creative Commons BY 4.0 or CC0.

[1] *A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)*

### 2.4.3. Will the data produced in the project be useable by third parties, in particular after the end of the project?

The intention of the project consortium is that data will be organised, curated and shared in a way that it will be understandable and usable by third parties after the end of the project. Although some qualitative data will not be made available for sharing for reasons of confidentiality (to ensure anonymity), we will ensure the openness of all other data wherever possible.

### 2.4.4. Will the provenance of the data be thoroughly documented using the appropriate standards?

All TIER2 datasets will maintain links with their raw or processed data that delimit the initiation of the data processes in the context of the project. Reused datasets support provenance via documenting the sources where they have aggregated content and the state of records at given time, the methods that they have used to process them and workflows that they have in place to curate, share and preserve them (e.g. history, version control, linked metadata etc). New datasets will include provenance information, where possible directly in the metadata.

### 2.4.5. Describe all relevant data quality assurance processes.

Depending on the activity that the datasets will be derived from, appropriate data quality assurance processes will be followed:

- Setting up a scientific and technical committee to perform internal project peer review of results, e.g. consisting of by at least one consortium or advisory board member and the project coordinators.
- Data conforming to format specifications
- Use of tools for automatic checks to validate content
- Code review of data analysis code
- Consistency verified with data models and standards, e.g. the procedures followed by the OpenAIRE Graph at the technical level: https://graph.openaire.eu/about#architecture.

# 3. Allocation of resources

## 3.1. What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?

FAIR metadata publishing and archiving is covered by the repositories, while storage, back up and security during research is supported by means of institutional infrastructure resources.

## 3.2. How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

All costs for providing FAIR data in TIER2, as explained in the FAIR section of this document, have been incorporated in the overall budget of the project, following the Horizon Europe grant eligibility criteria. The Financial Coordination has already incorporated in its financial distribution the human capacity of data managers needed throughout the project as well as the infrastructure and services that enable their effective operation.

## 3.3. Who will be responsible for data management in your project?

The Task leaders will be responsible for managing the datasets that each will generate, collect or reuse. Below is an indication of responsibilities' allocation at this initial phase of the project:

**Table 2:** *Data Management Coordination Responsibilities.*

| Task | Name of Activity | Name of Dataset | Description of Dataset | Data Management coordination |
|---|---|---|---|---|
| 1.3 | Auto-Ethnography | Reproducibility Diaries | Diary entries written quarterly by five TIER2 consortium members concerning their thoughts, ideas and perspectives in relation to reproducibility issues, both in TIER2 and beyond. | AU |
| 2.3 | Development of the Reproducibility Hub | OpenAIRE Graph Dump | OpenAIRE Graph is an open resource that aggregates a collection of research data properties (metadata, links) available within the OpenAIRE Open Science infrastructure for funders, organizations, researchers, research communities and publishers to interlink information by using a semantic graph database approach. | VUmc |
| 2.3 | Development of the Reproducibility Hub | FAIRsharing | FAIRsharing is a curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies, across all disciplines. It enables the FAIR Principles by promoting the value and | VUmc |

| | | | | |
|---|---|---|---|---|
| | | | use of data and metadata standards, and their use by databases. | |
| 3.2 | Subtask: Integrative review of reproducibility and qualitative research | Integrative Review Materials | A list/spreadsheet of DOIs of reviewed literature; a spreadsheet of extracted data from and process decisions about reviewed literature | KNOW |
| 3.3 | Synthesis and recommendations | Recommenda-tions Delphi Process | Interviews/transcripts from workshops; survey data from Delphi; Spreadsheets reporting (1) survey results; (2) first phase recommendations; (3) second phase recommendations; and (4) third phase recommendations. | KNOW |
| 4.1 | Future studies | Workshop results | Workshop transcripts; completed miro boards; analysed/processed data | VUmc |
| 4.2 | Pilot development | GESIS data | GESIS data that are relevant for the reproducibility studies of TIER2 will be reused. Examples are: (1) TweetsKB; a public RDF corpus of anonymized data for a large collection of annotated tweets. The dataset currently contains data for nearly 3.0 billion tweets, spanning more than 9 years (February 2013 - August 2022), and (2) ClaimsKG; a structured database which serves as a registry of claims. It provides an entry point for researchers to discover claims and involved entities, also providing links to fact-checking sites and their results | GESIS |

## 3.4. How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

The consortium follows a federated approach, where each partner is responsible for gathering the data, ensuring security, and long-term preservation via the used repositories, i.e. Zenodo, FAIRsharing, OSF.

# 4. Data security

## 4.1. What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Data stored in One Drive institutional storage are backed up incrementally or regularly as part of data security measures followed by the respective institutional providers, mainly Microsoft. For an extra layer of security, TIER2 partners will follow the 3-2-1 back up rule where 3 copies of the data (production data and 2 backup copies) are stored on two different media (disk and tape) with one copy off-site for disaster recovery. The same applies for Zenodo, FAIRsharing and GESIS that generate backups of their live content and keep it in their disk storage capabilities. For secure access to the data, servers will be protected by passwords and firewalls. Data that appear to have privacy constraints and applicable ethical norms will be anonymised and no raw data will be openly accessible. Instead, the processed and analysed data will be shared through deliverables and the metadata will be made openly available.

## 4.2. Will the data be safely stored in trusted repositories for long term preservation and curation?

In most of the times, the data will be preserved along with their metadata records in the trusted repositories of Zenodo, FAIRsharing and OSF that the project will be using.

# 5. Ethics

## 5.1. Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

For privacy reasons and because some of the data might contain sensitive and personal information about the project members and external participants, all data will be securely stored and is not openly accessible. The findings of the processed data will be included in reports, such as a summary and analysis of the reproducibility diaries data in the self-reflection report (D1.3).

## 5.2. Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?

The consortium partners will provide external participants with consent forms before collecting their input. The forms will be made available as project publications and reference data sharing and long-term preservation for future studies. A first example of a consent form created for the needs of Participation Information Document can be found at: https://osf.io/c7ka6. The document explicitly refers to “I understand that the data will be archived in repository of the Open Science Framework platform for ten years for scientific integrity. I understand that other researchers will have access to this data only if they agree to preserve the confidentiality of the data.”
A template for consent forms to be used across the project is in development.

# 6. Dataset Descriptions

In this section, we provide more detail about identified datasets that we already know TIER2 will collect, generate and/or re-use. Some of the datasets refer to project activities that have not started, and they might change significantly in the future. Similarly, more datasets are expected to be described in the next iteration of the DMP (M18).

## 6.1.Reused Datasets

**Table 3a:** *Reused Datasets: Summary.*

| No | Name | Description | Type | Format | Origin / Provenance | Used Software | Data Utility |
|---|---|---|---|---|---|---|---|
| **R-1** | OpenAIRE Graph Dump | Re-using contextualised data from the OpenAIRE Graph to facilitate T2.3 in creating graphs based on indicators for reproducibility. | Derived or compiled: The data are a list of DOIs gathered and reviewed for an integrative review of how reproducibility/replicability are conceived in relation to qualitative research, as well as which open science practices are discussed in relation to supporting reproducibility of qualitative research. | JSON, XML, PDF | Metadata are harvested from trusted sources and all links are kept with the original resource. New links created upon curation of the data in the Graph are also kept for every iteration of the algorithm during monthly updates. History of those changes and enhancements are made available to resource providers using the OpenAIRE PROVIDE service (https://provide.openaire.eu/home). | The data will be the input of the analysis algorithm that will be developed in the context of the project to exploit the data and offer visualisations in the form of graphs. | • Researchers<br>• Research communities<br>• Decision makers<br>• Economy<br><br>OpenAIRE data contain rich information about science and its evolution, especially on the Open Science realm. There are immediate and meso-/ long - term potentials from exploiting OpenAIRE Graph reused data from the perspective of reproducibility. They both are able to positively affect the research sector at different pace and levels:<br><br>Researchers and research communities can use the data from the dump, in the same way that TIER2 is getting them, and they can build on top of them and analyse them based on their own scientific interests and objectives.<br><br>Decision and policy makers can embed those data in their scientific ecosystems to influence and enhance intelligence policies for science.<br><br>The economy will eventually flourish by keeping track of the evolution of reproducibility through periodic exploitation and enhancement of those data and their reproducibility algorithms, and by continuously healing identified gaps that mitigate reproducibility risks. |

| No | Name | Description | Type | Format | Origin / Provenance | Used Software | Data Utility |
|---|---|---|---|---|---|---|---|
| **R-2** | FAIRsharing dataset | FAIRsharing is a curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies, across all disciplines. FAIRsharing guides consumers to discover, select and use these resources with confidence, and producers to make their resource more discoverable, more widely adopted and cited. FAIRsharing enables the FAIR Principles by promoting the value and use of data and metadata standards, and their use by databases. FAIRsharing is available via both human- and machine-accessible options. Access to the FAIRsharing metadata for computational purposes is described here and is covered by a CC-BY-SA 4.0 licence (see also here). | "Derived or compiled: Across the research disciplines there are thousands of standards and several thousands of databases, designed to assist the virtuous data cycle, from collection to annotation, through preservation and publication to subsequent sharing and reuse. As consumers of these standards and databases, it is often difficult to know which resources are the most relevant for your specific domain and needs. As producers, you want to be sure your standard or database is findable by prospective users, and recommended in data policies by funders, journals and other organisations. With our growing and interlinked content, functionalities and endorsements, FAIRsharing is the most comprehensive informative and educational resource of standards, databases and policies. FAIRsharing is a web-based, | JSON | Each record is manually curated based on publicly-available information about the standard, database or policy it describes. This manual curation is done by in-house curators and community volunteers. These volunteers are further divided into maintainers (who are responsible for the resource being described) and community champions (who may edit records across their research domain of interest). | The data will be the input of the analysis algorithm that will be developed in the context of the project to exploit the data and offer visualisations in the form of graphs. | • Researchers<br>• Research communities<br>• Decision makers<br>• Other<br><br>FAIRsharing is a community-driven resource with users and collaborators across all disciplines. We work together with our stakeholders to enable the FAIR Principles by promoting the value and the use of standards, databases and policies. These stakeholders within TIER2 include:<br><br>Developers & curators of resources and tools: Integration of their resources with ARGOS, the Reproducibility checklist, and/or the FAIRsharing collections that will be produced.<br><br>Journal publishers, funders and other policymakers: TIER2, especially WP4, will be working closely with policymakers on a proposed “Reproducibility Checklist, policy and practices” intervention. For these policymakers, FAIRsharing can be used to understand the landscape of resources relevant to their implementors, and also as a method of transmitting their requirements to them. For funders, this includes improvements to policies and creation/curation of the relevant FAIRsharing Collections. For publishers, this primarily involves the creation of the reproducibility checklist. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | searchable portal of three interlinked registries, containing both in-house and crowd-sourced manually curated descriptions of standards, databases and data policies, combined with an integrated view across all three types of resource. | | | | Researchers, research data facilitators, librarians, trainers: For them, the utility of FAIRsharing will be in: FAIRsharing's contribution to the Reproducibility checklist, creation/curation of FAIRsharing Collections of standards and databases, and connectivity of FAIRsharing to ARGOS via T2.3. |
| **No** | **Name** | **Description** | **Type** | **Format** | **Origin / Provenance** | **Used Software** | **Data Utility** |
| **R-3** | GESIS data | GESIS data that are relevant for the reproducibility studies of TIER2 will be reused. Examples are:<br>- TweetsKB: a public RDF corpus of anonymized data for a large collection of annotated tweets. The dataset currently contains data for nearly 3.0 billion tweets, spanning more than 9 years (February 2013 - August 2022)<br>- ClaimsKG: a structured database which serves as a registry of claims. It provides an entry point for researchers to discover claims and involved entities, also providing links to fact-checking sites and their results | Derived or compiled: The data will be a collection of GESIS archived artifacts, compiled for the needs of TIER2 project. | Varying: .gz, .ttl | Other studies that have been the occasion of the collection, curation and sharing of those datasets. Some previous versions of the datasets might be available on other repositories. Provenance information and links to past versions are available upon their download. | | |

**Table 3b:** *Reused Datasets: FAIR.*

| No | PID | Metadata | Keywords | Vocabularies | Access | Repository | Licenses | Size | Data Manager |
|---|---|---|---|---|---|---|---|---|---|
| **R-1** | The full list of the PID types that the OpenAIRE Graph collects can be found here: https://api.openaire.eu/vocabularies/dnet:pid_types. | Metadata about the OpenAIRE Research Graph is searchable via Zenodo and OpenAIRE itself (via explore.openaire.eu). They are offered according to the OpenAIRE metadata format: a. https://zenodo.org/record/4723403; b. https://doi.org/10.5281/zenodo.3974225. | Knowledge Graphs; SKGs; Scholarly Communication; Open Science; EOSC | The full list of OpenAIRE vocabularies can be found here: https://api.openaire.eu/vocabularies/. | Open Access | Zenodo (zenodo.org) | Creative Commons Attribution 4.0 | 240GB | Elli Papadopoulou (orcid:0000-0002-0893-8509) |
| **R-2** | • Data identifiers - DOI<br>• Researchers identifiers - ORCIDs<br>• Projects identifiers-ROR | The metadata are provided according to the FAIRsharing schema: https://zenodo.org/record/6884446 | Standardisation; Database | The subject and domain ontologies draw upon over 50 community-developed ontologies across a variety of domains for tagging, incl. the NCBI Taxonomy for taxonomic scope, where appropriate. List of used vocabularies: https://github.com/FAIRsharing/subject-ontology (see also https://doi.org/10.25504/FAIRsharing.b1xD9f); | | FAIRsharing registries | | | Allyson Lister (orcid:0000-0002-7702-4495) |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | https://github.com/FAIRsharing/domain-ontology (see also https://doi.org/10.25504/FAIRsharing.FSIfv8); NCBI Taxonomy (https://doi.org/10.25504/FAIRsharing.fj07xj). | | | | | |
| **No** | **PID** | **Metadata** | **Keywords** | **Vocabularies** | **Access** | **Repository** | **Licenses** | **Size** | **Data Manager** |
| **R-3** | • Data identifiers - DOI | Descriptive metadata, provided following the schema. | Social Data; Social Media; Social Studies | GESIS Thesaurus of scientific domains for descriptive metadata | Various access schemes, based on the selected dataset, e.g. open, shared. | GESIS - Leibniz-Institute for the Social Sciences | Various licenses; some might be restricting reuse for non-commercial research. | | Hajira Jabeen (orcid:0000-0003-1476-2121) |

## 6.2.New Datasets

**Table 4a:** New Datasets: Summary.

| No | Name | Description | Type | Format | Origin / Provenance | Methods, incl. Reproducibility | Used Software | Data Utility |
|---|---|---|---|---|---|---|---|---|
| **N-1** | **Reproducibility Diaries** | Diary entries written quarterly by five TIER2 consortium members concerning their thoughts, ideas and perspectives in relation to reproducibility issues, both in TIER2 and beyond. | Observational: The dataset offers a collection of personal thoughts and experiences from project partners' activities. It captures how project partners navigate their everyday life in different research endeavours, how aware they are of the elements that mitigate reproducibility risks, how they engage in practicing reproducibility and they operationalise this knowledge in discussion with others. | AbiWord Document | Digital, written diary entries in Word/pdf format, by project members of the TIER2 consortium. An overview of data provenance will be provided, including dates and format of the diary entries. In addition, the final deliverable reporting on the data, will include detailed descriptions of the way in which the data were processed and analysed. | The data consist of diary entries written by five project members on a quarterly basis throughout the project duration. Based on a flexible and open format, members write about their ideas, concerns, practices and discussions related to reproducibility issues, both related to the TIER2 project as well as beyond. Entries are written individually though they will also reflect on group discussions and interactions within and beyond the consortium. The dataset will be accompanied by readme files. Negative results might be shared as part of project members' experience. For more details on methodology, please consult project deliverable 1.3. | https://lumivero.com/products/nvivo/#:~:text=What%20is%20NVivo%3F,from%20their%20qualitative%20data%20faster. | • Researchers<br>• Research communities<br>• Decision makers<br>The reflections derived from the data may support researchers, research policymakers and their communities to optimally organise and coordinate international, multidisciplinary research projects in terms of reproducibility issues. The data will shed light on the practices that researchers could employ to foster reproducibility and the kind of concerns or obstacles they face when trying to implement these practices. This includes potential disciplinary or organisational barriers towards reproducibility and will henceforth inform future policymakers, researchers and their communities to smoothen the path towards increased reproducibility standards. |
| **N-2** | **Futures Studies - Workshop results** | Workshop transcripts; completed miro boards; analysed/processed data | Observational | Acrobat PDF 1.0 - Portable Document Format | Primary data from audio recordings of the workshops. | Inductive content analysis; Codebooks | | • Researchers<br>• Research communities<br>• Decision makers<br>• The public |

| No | Name | Description | Type | Format | Origin / Provenance | Methods, incl. Reproducibility | Used Software | Data Utility |
|---|---|---|---|---|---|---|---|---|
| **N-3** | **Integrative Review Materials** | A list/spreadsheet of DOIs of reviewed literature; a spreadsheet of extracted data from and process decisions about reviewed literature | Derived or compiled: The data are a list of DOIs gathered and reviewed for an integrative review of how reproducibility/replicability are conceived in relation to qualitative research, as well as which open science practices are discussed in relation to supporting reproducibility of qualitative research. | Basic Excel spreadsheet | | This dataset will be generated using the method for an integrative literature review, as described here: http://link.springer.com/10.1007/978-3-030-37504-1. It will be accompanied by readme files. | The resulting dataset will have been derived from conducting an integrative review using the SyRF platform: https://syrf.org.uk/ | • Researchers<br>• Research communities<br>• Education<br>This dataset will be of interest to researchers studying the reproducibility of qualitative research and educators interested in teaching and supporting reproducibility of qualitative research. |
| **N-4** | **Recommendations Delphi Process** | Interviews/transcripts from workshops; survey data from Delphi; Spreadsheets reporting (1) survey results; (2) first phase recommendations; (3) second phase recommendations; and (4) third phase recommendations. | Observational: This dataset includes qualitative data that consist of recorded videos, transcripts, and step-wise data charting that document discussions, debates and brainstorming of science policy recommendations to support reproducibility of research gathered during a multi-phased co-creative Delphi process, in response to the cumulative findings of the Tier2 project. | Basic Excel spreadsheet | | This dataset will be created using a co-creative modified Delphi process, as described here by the researchers: https://royalsocietypublishing.org/doi/10.1098/rsos.221460. It will be accompanied by readme files. | | • Researchers<br>• Research communities<br>• Decision makers<br>• Education<br>The data contained herein may be instructive and/or useful to researchers, research communities, policy-makers and institutional leaders at higher education institutions interested in the reproducibility of research and how best to foster it. |

**Table 4b:** *New Datasets: FAIR.*

| No | PIDs | Metadata | Keywords | Vocabularies | Access | Repository | Licenses | Size | Data Manager |
|---|---|---|---|---|---|---|---|---|---|
| **N-1** | • Data identifiers – DOI<br>• Researchers' identifiers – ORCIDs<br>• Projects identifiers- Cordis | Descriptive according to DataCite. | Diary Entry; Reproducibility; Multidisciplinary Research; Organisational Challenges | Key words, abstract and project information will be drafted in line with other project outputs and use the vocabulary commonly used in the main project deliverables. The vocabulary used will specifically be based on the concepts and terminology described in Milestone 3.1 of the project. | Given the sensitive and personal nature of the data, not all data will be shared. However, the processed data that can be securely shared will be openly available. | Zenodo (zenodo.org) | Creative Commons Attribution 4.0 | 10MB | Serge P. J. M. Horbach (orcid:0000-0003-0406-6261) |
| **N-2** | Data identifiers - DOIs | Descriptive according to the OSF framework. | | | Open Access | Open Science Framework (osf.io) | Creative Commons Attribution 4.0 | 1GB | Joeri Tijdink (orcid:0000-0002-1826-2274) |
| **N-3** | • Data identifiers – DOI<br>• Researchers' identifiers – ORCIDs<br>• Projects identifiers- Cordis | Descriptive according to DataCite. | Reproducibility; Open Science | | Open Access | Zenodo (zenodo.org) | Creative Commons Attribution 4.0 | 10MB | Nicki Lisa Cole (orcid:0000-0002-6034-533X) |
| **N-4** | • Data identifiers – DOI<br>• Researchers' identifiers – ORCIDs<br>Projects identifiers- Cordis | The dataset will follow the guidance on qualitative data sharing: https://qdr.syr.edu/guidance/managing/preparing-data | Reproducibility; Open Science; Qualitative data | | Open Access | Zenodo (zenodo.org) | Creative Commons Attribution Share-Alike 4.0 | 10MB | Nicki Lisa Cole (orcid:0000-0002-6034-533X) |

# References

[1] Jahn, Najko, Laakso, Mikael, Lazzeri, Emma, & McQuilton, Peter. (2023, March 21). Study on the readiness of research data and literature repositories to facilitate compliance with the Open Science Horizon Europe MGA requirements (Version Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.7728016

[2] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

# Annex

Metadata of badging reproducible practices as suggested by NISO (https://groups.niso.org/higherlogic/ws/public/download/24810/RP-31-2021_Reproducibility_Badging_and_Definitions.pdf):

- Version of the schema or specification
- Issuing organization
- Badge type
- Badge definition
- Paper DOI
- Issuing date
- References (linked DOIs to artifacts)
- Review criteria URI (for the ROR badge)
- Optional: validation hash or cryptographic key

Funded by
the European Union