REPAiR

REsource Management in Peri-urban AReas: Going Beyond Urban Metabolism

D8.4 Draft Research Data Management Plan

Version 4.0

Author(s): Rusnė Šilerytė (TUD)
Alexander Wandl (TUD)
Jasmin Böhmer (TUD)
Max Bohnet (GGR)
Jens-Martin Gutsche (GGR)
Sue Ellen Taelman (UG)
Maria Cerreta (UNINA)
Viktor Varjú (RKI)
Konrad L. Czapiewski (IGiPZ)
Andreas Obersteg (HCU)

Grant Agreement No.:	688920
Programme call:	H2020-WASTE-2015-two-stage
Type of action:	RIA – Research & Innovation Action
Project Start Date:	01-09-2016
Duration:	48 months
Deliverable Lead Beneficiary:	TUD
Dissemination Level:	PU
Contact of responsible author:	r.sileryte@tudelft.nl

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 688920.

Dissemination level:
PU = Public
CO = Confidential, only for members of the consortium (including the Commission Services)

Change Control

V	Date	Author	Organisation	Description/ Comments
1.0	01.02.2017	Rusnė Šilerytė	TUD	First draft version
1.1	08.02.2017	Rusnė Šilerytė, Alexander Wandl	TUD	Draft version ready for collaborative writing
1.2	24.02.2017	Andreas Obersteg, Sue Ellen Taelman, Viktor Varju, Peter Eder, Annie Attademo, Maria Cerreta, Libera Amenta, Konrad L. Czapiewski	HCU, UG, RKI, JRC, UNINA, IGiPZ	Comments and edits on the draft version
2.1	27.02.2017	Jasmin Bohmer	TUD	Imput from 4TU.Centre for Research Data
2.2	27.02.2017	Alexander Wandl	TUD	Final version for submission
3.0	28.02.2017	Rusnė Šilerytė	TUD	Final layout and editting
4.0	28.06.2018	Rusnė Šilerytė, Max Bohnet, Jens-Martin Gutsche, Alexander Wandl, Jasmin Bohmer	TUD, GGR	Review and corrections after 18 project months

Acronyms and Abbreviations

4TU	4TU-cooperation of TU Delft, Eindhoven University of Technology, University of Twente and University of Wageningen
CE	Circular Economy
D	Deliverable
DMP	Data Management Plan
DOI	Digital Object Identifier
EC	European Commission
EU	European Union
GA	Grant Agreement
GDSE	Geodesign Decision Support Environment
Geo-Col	Geo-Col GIS and Collaborative Planning, Amsterdam
GGR	Gertz Gutsche Rümenapp - Stadtentwicklung und Mobilität, Hamburg
GIS	Geographic Information System
HCU	HafenCity University, Hamburg
LCA	Life Cycle Assessment
MFA	Material Flow Analysis
IGiPZ	Institute of Geography and Spatial Organisation, Polish Academy of Sciences, Warsaw
INSPIRE	Infrastructure for Spatial Information in Europe
OSF	Open Science Foundation
PULL	Peri-Urban Living Labs
RKI	Institute for Regional Studies, Centre for Economic and Regional Studies of the Hungarian Academy of Sciences, Pécs
SQL	Structured Query Language
TUD	Technical University of Delft, Delft
UG	University of Ghent, Ghent
UNINA	University of Naples Federico II, Naples
WMS	Web Map Service
WFS	Web Feature Service
WP	Work Package

Contents

Publishable Summary

This deliverable describes the preliminary version of the Data Management Plan (DMP) for the REPAiR project. The DMP provides a draft summary of the main elements of the data management policy that will be used throughout the REPAiR project by the project partners, with regard to all the data that will be generated and used by the project.

The DMP is a living document in which information can be made available on a finer level of granularity through updates as the implementation of the project progresses and when significant changes occur.

The format of the plan follows the Horizon 2020 template[1] and is supported by the 4TU.Centre for Research Data[2].

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] http://researchdata.4tu.nl/en/planning-research/data-management-plan/

1 The DMP in the overall REPAiR project approach

1.1 What is a Data Management Plan?

The Data Management Plan (DMP) is a document that describes the data management starting from its collection, including its processing and handling during the REPAiR project and, finally, its later archiving and dissemination. It is a living document in which information can be made available on a finer level of granularity through updates as the implementation of the project progresses and when significant changes occur.

The document helps project participants to determine how the data can be managed efficiently and effectively and reduce the risk of data loss and conflicts. Ethical issues and data security are also briefly discussed within the document. Finally, the plan ensures consistent resource and budgetary planning for data management related costs.

DMP lists the types and specifications of data that is and will be collected, generated, processed or generally, used, during the project. The specifications include detailed descriptions of data handling methodologies and used standards both within the project team and outside. Moreover, the DMP includes information on handling the research data both during the project and after it is finished.

According to the guidelines provided by the Horizon 2020, the project data must be FAIR (Findable, Accessible, Interoperable and Reusable) as much as it is possible, except when there are substantial reasons to keep the data confidential. Among other aspects, the DMP also describes the possible levels of data openness within the REPAiR project, the differences in their handling and compliance with the Intellectual Property Rights (IPR).

The format of the plan follows the Horizon 2020 template[1] and is supported by the 4TU.Centre for Research Data[2].

1.2 REPAiR project specific data management aspects

The core objective of REPAiR project is to provide local and regional authorities with an innovative transdisciplinary open source Geodesign Decision Support Environment (GDSE) developed and implemented in Peri-Urban Living Labs (PULLs) in six metropolitan areas. The GDSE allows creating integrated, place-based eco-innovative spatial development strategies aiming at a quantitative reduction of waste flows in the strategic interface of peri-urban areas. These strategies will promote

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] http://researchdata.4tu.nl/en/planning-research/data-management-plan/

the use of waste as a resource, thus support the ongoing initiatives of the European Commission towards establishing a strong Circular Economy.

The GDSE consists of three main components: hardware, software and process-ware. This means that there will be multiple types of data collected and used throughout the project. The main purposes of data can be divided into the two following groups: 1) data used during the research process internally by the project partners; and 2) data used during PULLs as an integral part of the decision support environment. The first group consists of data that is used to determine the current status quo of the system under investigation, prepare the models to be used within the software and, finally, analyse feedback after the PULLs. This group has higher variety of data types, contains more raw and straight-from-the-source data but also derived and designed data, which can be provided in various formats. The second group requires strict data structure, unified formats and units, clear semantics and provenance.

Based on the above explained grouping, the data can also be divided into:

1. Digital Objects (e.g. text files and audio files, images, posters, also websites, programmed modules, etc.);
2. Databases (structured collections of records stored in a computer system).

Specific data groups and their utility are discussed in Section 2: Data Collection and Generation.

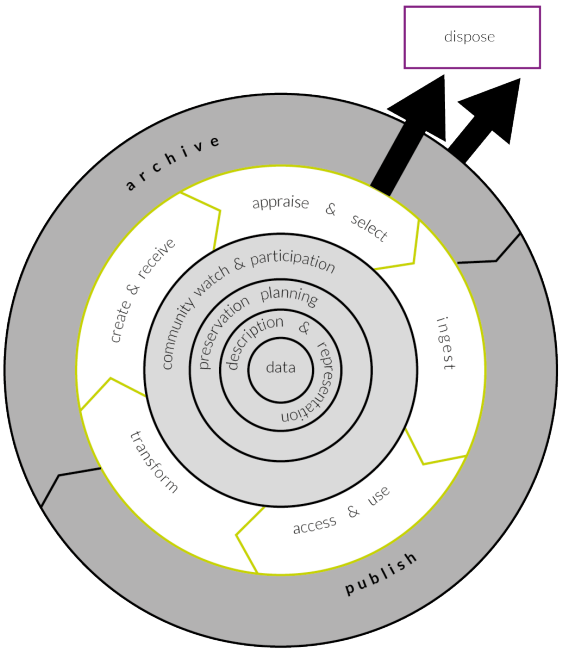


Figure 1.1. *Different stages of the Data Life Cycle, based on the Data Life Cycle developed by the Digital Curation Center*

1.3 The overall structure of the Data management in REPAiR

"The Data Life Cycle (DLC) provides a high level overview of the stages involved in successful management and preservation of data for use and reuse. Multiple versions of a DLC exist with differences attributable to variation in practices across domains or communities." (Le Franc, 2017)[3]

The DMP is based on the Data Life Cycle and its different stages and dimensions. There are multiple DLCs available from different communities, however the one that fits closest with the DLC stages of REPAiR has been provided by the Digital Curation Center[4] (Figure 1.1). The cycle should be read in two dimensions - there is a set of sequential actions and procedures that follow them throughout the full cycle. The further sections of DMP will refer back to the stages of DLC.

In the following paragraphs first the data management structure during the project period is presented and thereafter the long term data management strategy.

1.4 Data management during the project

The main data management processes, roles and storage elements within REPAiR project are depicted in Figure 1.2.

There are 3 data management roles within REPAiR:

Data Captain - each case study has dedicated Data Captains who have the responsibility to deliver their data to the project according to the agreed rules. Data Captains are responsible to verify that the data they are providing is sufficiently described using metadata, complies to the agreed rules described in the DMP and meets the defined deadlines.

REPAiR Researcher - everyone who belongs to the REPAiR consortium. Each researcher is able to provide data which is either self-collected, generated using other data or reused from the other sources. The researcher who is providing the data is responsible to describe it using metadata and deliver to the corresponding Data Captain. Each researcher within REPAiR is also able to access and use data that has been delivered to the consortium, unless the data is has specific access restrictions.

User / 3rd party researcher - anyone outside of REPAiR consortium is able to access, make use of, give feedback for and cite the data that has been marked as `public`.

User roles will be defined with the corresponding rights to read and write data. For sensitive data, e.g. personalised interview recordings, private components will be set up, where only Data Captains will have access rights.

For the REPAiR project the OpenScienceFramework (OSF.io)[5] will be used as a platform for uploading data and tagging the necessary metadata. The Open Science Framework (OSF) is partly storage and access interface of research materials, partly version control system, and partly a platform for collaboration (Figure 1.2). Web-based project management reduces the likelihood of losing study materials due to computer malfunction, changing personnel, or just forgetting. OSF is an open source tool that provides the features required by the REPAiR project:

[3] https://www.slideshare.net/EUDAT/the-data-lifecycle-eudat-summer-school-yann-le-franc
[4] www.dcc.ac.uk , http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf
[5] https://osf.io/

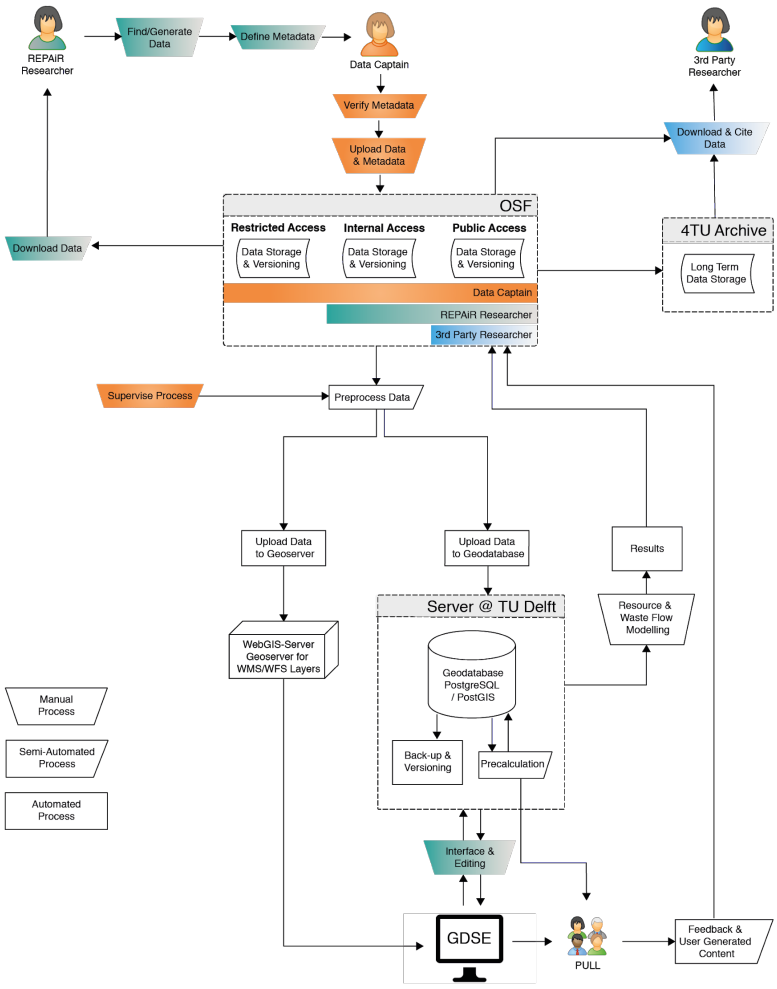


Figure 1.2. *Main elements of the data management within REPAiR during the project.*

- It provides a secure cloud storage;
- It respects differentiated access rights;
- It enables a direct exchange of data to the modelling server hosted at TU Delft;
- It ensures automated versioning of data;
- All data can be cited in scientific publications via permalinks.

The usage of the OSF web page and tools will be organised as follows:

- For all researchers within the REPAiR project the OSF web page will be the first and central online user interface to be used to retrieve project data.
- Each of the six PULL-regions will have an own folder (called `Component` in OSF) to organise region-specific data.

- Intermediate and final results of the models and simulations (being processed on the modelling server at TU Delft) will be provided in the GDSE for users with permissions for a case study and will also be provided on the OSF platform to all REPAiR researchers in dedicated folders.
- Working files such as project deliverables, reports, conceptual schemes, etc. will not be kept in OSF but in a dedicated folder on Google Drive[6] that will be shared among all consortium members. All consortium members will be granted permissions to access, modify and share files from Google Drive, however, only a small group of people will be granted these rights on OSF. Therefore all finite and sensitive information must to be stored on OSF.
- All programme code will be hosted on publicly accessible GitHub[7] page[8], which allows versioning of all code. The GitHub page will be integrated into the OSF platform through an add-on.
- Templates for data upload will be stored in a Common folder and described in a separate living document[9]
- Supporting material for the deliverables and other kinds of publications will be stored in public folders.
- For all uploaded data the DMP defines, which metadata have to be provided as in the following sections. Data delivery and metadata rules are also described in a separate living document [10]
- Data upload on Geoserver will be conducted through OSF ownCloud add-on, instruction for upload will provided in a separate living document[11]
- For the long term Storage, the REPAiR will use the 4TU.Datacenter, which guarantees a long term availability of the project results. The data will transferred directly from OSF platform to the 4TU archive based on the data preservation tag that is given while filling in the metadata file.

Alongside the OSF platform a PostgreSQL geodatabase with PostGIS spatial extension is set up in a TU Delft server. The server is needed for the direct data access by the GDSE web application, data processing and modelling. The PostGIS-Server also contains spatial databases for each case study with large datasets directly imported from Open-Data Sources like *openstreetmap*, the *European Environmental Agency*, etc. These Datasets can be used in modelling certain indicators or in map representations in the geoserver. Due to the size of the datasets only a direct bulk import into the PostGIS Database is feasible and no copy of this data is stored on OSF. However, the metadata of these datasets will be stored on OSF.

1.5 Long-term data management

After the research is finished, the data that does not have privacy restrictions (including necessary documentation, metadata, code, consent form, software, etc.) will be stored and made publicly available in the 4TU.Centre for Research Data

[6] https://www.google.com/drive/
[7] https://github.com/
[8] https://github.com/MaxBo/REPAiR-Web
[9] https://docs.google.com/document/d/1HvTsxFcAT_PL3jnUttRtiJL1MV9XGoxHwZdjY47zZAM/edit?usp=sharing
[10] https://docs.google.com/document/d/1SLm9Ac169V-7nTOlUBrYS0SBZpTSh7bSUGPEjNAlpOA/edit?usp=sharing
[11] https://docs.google.com/document/d/1ZaLGukHtBrWrWT-6yJEJyawOYA_4B1zW-uAMQdjht2I/edit?usp=sharing

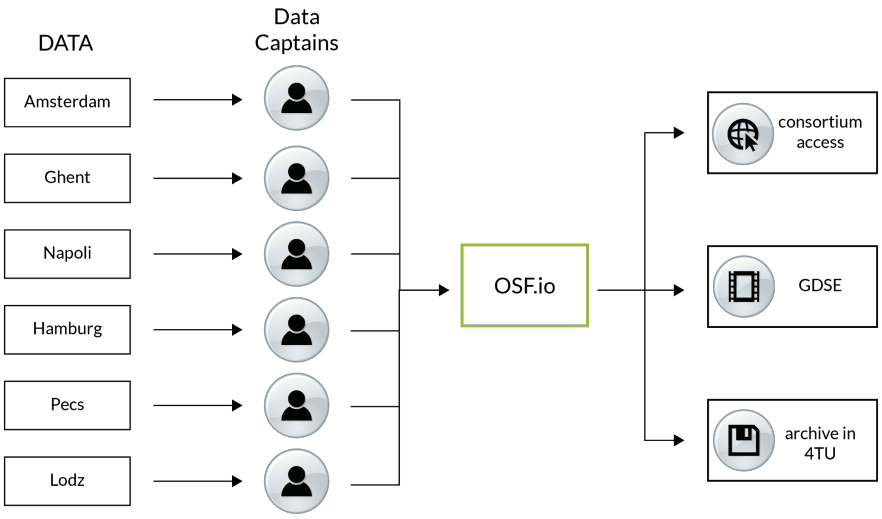


Figure 1.3. *Role of the OSF platform in REPAiR Data Management*

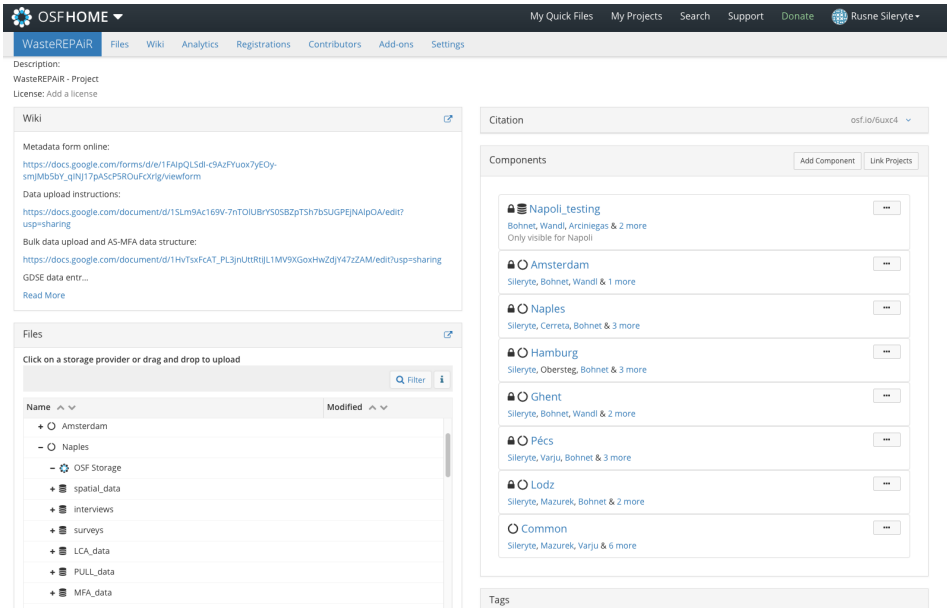


Figure 1.4. *REPAiR data repository on OSF platform*

archive under the Deposit License agreement. By doing so, 4TU.Centre for Research Data is granted a non-exclusive licence to store the data and make them available to third parties.The General Terms of Use will apply for the data (re)users. In short, these terms specify that, when reusing the data, they will clearly state the name(s) of the original author(s) and that the data will not be used for commercial purposes.

4TU.Centre for Research Data archive is a long term archive with a Data Seal of Approval. The 4TU.Centre for Research Data will keep the data available for at least 15 years in an open and - if needed - closed data archive. This complies with The Netherlands Code of Conduct for Academic Practice according to which ”raw research data are stored for at least ten years. These data are made available to other

academic practitioners upon request, unless legal provisions dictate otherwise."

The 4TU.Centre for Research Data offers a solid service that enables findability, accessibility and interoperability. Current developments at 4TU are working towards improving the reusability of datasets. In order to provide all necessary information, every deposited REPAiR dataset will be accompanied by a general metadata file that among other aspects explains methodologies, software setups and experiment setups in detail.

The detailed data preservation plan of the 4TU.Center for Research data is available online[12] and is update regularly.

1.6 Allocation of Resources

The budget within REPAiR is allocated in two ways to make data management FAIR (findable, accessible, interoperable and reusable):

Long term storage. The costs for depositing data at the 4.TU Center for Research Data for 15 years are €4,50 per GB. In total about €6.000 are allocated for long term data storage. This cost is estimated to be sufficient for storing all preservable REPAiR research data of all case studies.

Journal open access costs. €30.000 are allocated for OpenAccess Publishing, cost for open access publishing vary significantly dependent on the scientific journal but this should be sufficient for at least 10 peer reviewed articles.

Key responsible for the overall data management is Alexander Wandl (TUD), in case of his absence Arjan van Timmeren (TUD) is responsible. Each case study has assigned one person responsible for data management within one's case study. The responsible person is called Data Captain who:

- is responsible for uploading all data related to their case study into OSF;
- verify that sufficient metadata is provided;
- ensure that data structure, naming, etc. comply to the DMP;
- manage data privacy.

Each Data Captain can have one or more Captain Assistants who help uploading data, collecting and writing metadata and generally preparing the files for the OSF. However, Data Captains are still held responsible for uploading the data on agreed time in agreed formats as described in D2.2 Data Delivery Plan. Table 1.1 lists all Data Captains and their Assistants within the project alongside with their contact details.

OSF permits differentiated access rights for its users (called Contributors) which can be defined per each component. The rights can be:

Read :

- can see and download the contents.

Read/Write :

- can create new components;

[12] http://researchdata.4tu.nl/fileadmin/editor_upload/pdf/Preservation_Plan/4TU.Preservation_Plan.pdf

- can add new data /delete old data;
- can edit Wiki.

Administrator :

- can add/delete Contributors;
- can register the project;
- can control privacy;
- can delete other components.

The osf.io data repository is administered by Max Bohnet (GGR) and Rusne Sileryte (TUD). Data Captains and their assistants will be granted **Read/Write** rights for their respective case study components and for those components that are common to all case studies. Anyone in the consortium may be granted **Read** rights upon request for the whole project. If a person who has any rights granted quits the project, he/she will be denied from further access. Each case study will have a dedicated component for storing confidential data that may only be accessed by the Data Captain and his/her Assisstant and Administrators. The Data Captains will be responsible to decide which data needs to be stored in such a component.

Case	Data Captain	Assistant Captain
Amsterdam	Alexander Wandl	Rusne Sileryte
Ghent	Sue Ellen Taelman	
Hamburg	Andreas Obersteg	Alessandro Arlati
Lodz	Damian Mazurek	
Naples	Maria Cerreta	Pasquale Inglese
Pecs	Viktor Varju	Tamás Szabó
Administrator	Max Bohnet	Rusne Sileryte

Table 1.1. *Data Captains and their Assistants within the project.*

2 Data Collection and Generation

The major groups of data used in the REPAiR project are as ellaborated in the following sections.

2.1 Spatial (geographical) data

Purpose. Use of geographical data is mostly related with the following project tasks:

- T3.1 Spatial Analysis, where spatial data is used for the description of system's physical and administrative borders, infrastructures, land-use, typologies, planning specifications.
- T3.2 Material Flow Analysis, where spatial data is used to extend MFA with particular geographical knowledge.
- T3.3 Surveys of households and companies, where spatial data is used to relate socio-cultural and socio-economic factors with particular geographical contexts.
- T4.4 Generating Impact Models, where spatial data is used to spatialise estimated impacts.
- T5.2 and T5.4 Carrying out PULLs, where spatial data is used to introduce PULL participants to series of maps representing the status quo of the system, possible changes and their impacts.

Types & Formats. SHP files accompanied by SLD layers

Origin. Mostly reused from available data sources such as national and regional databases and openly available Volunteered Geographical Information; or generated by merging the available sources with non-spatial types of data.

Data utility. After the project is finished, the generated spatial data will be useful for other researchers who aim to connect resource flow patterns and their impacts with the particular geographical contexts.

2.2 Material Flow Analysis (MFA) data

Purpose. Use of MFA data is mostly related with the following project tasks:

- T3.2 Material Flow Analysis, where data is used to carry out the analysis.
- T4.3 Generating evaluation models, where MFA data is used as basic input for the evaluation models.

- T4.4 Generating impact models, where MFA data is used to represent the status quo.
- T5.5 Documentation of eco-innovative solutions, where MFA data is used to represent status quo that the solutions are going to modify.

Types & Formats. TSV files prepared according to the available templates.

Origin. Mostly international, national and regional databases.

Data utility. Data can be reused by anyone interested in Material Flow Analysis of the selected keyflows.

2.3 Life Cycle Assessment Data

Purpose. Use of LCA data is mostly related with the following project tasks:

- T4.2 Aggregation of sustainability indicators in the assessment framework
- T4.3 Generating evaluation models
- T4.4 Generating impact models

All the listed tasks rely on LCA data as basis for impact assessment in order to compare eco-innovative solutions.

Types & Formats. XLSX

Origin. Collected from global, national and regional databases, reports and provided directly by the project partners.

Data utility. If well documented and structured, when made available publicly, data can be reused by anyone conducting an LCA that involves the same processes as those covered by REPAiR.

2.4 Data generated during PULLs

Purpose. Use of data generated during PULLs is mostly related with the following project tasks:

- Task 5.5 Documentation of eco-innovative solutions;
- Task 6.2 Development of decision models for all case studies;
- Task 6.3 Implementation of decision (support) models;
- Task 6.4 Integration of the decision models and the geo-design decision support environment;
- Task 7.3 Organising knowledge transfer events as part of the PULLs in case study areas;
- Task 7.5 Creating an online handbook of transferable solutions and methods to facilitate transfer.

Data generated during PULLs will serve not only the specific case studies where it is generated but also the knowledge exchange between all of the case studies.

Types & Formats. Images, text, drawings, schemes, reports

Origin. Generated during PULL workshops

Data utility. The final catalogue of transferable solutions can be reused by anyone looking to adapt circular economy strategies in specific areas.

2.5 Software Code

Purpose. Software code is also considered to be part of the project data management as it needs consistent versioning, accessibility, collaborative work and documentation. It is, however, treated differently than all the rest of the project data. The main purpose of the project code is to enable and support the following tasks:

- T 2.1 Developing the five modules of the GDSE;
- T 2.2 Testing and Implementing the GDSE in the PULLs;
- T 2.3 Documentation of GDSE.

Types & Formats. Mostly Python and JavaScript files and their supporting file types.

Origin. The code is written by the project members in WP2. Some parts of the code may be reused from other open source projects and repositories.

Data utility. The finished code can be reused in parts for specific tasks or as a whole web application to be adapted to other case studies beyond the scope of REPAiR.

2.6 Interviews

Purpose. Data collecting during interviews will mostly serve the following tasks:

- T 6.1 Analysis of the decision making landscape in the case study areas;
- T 6.2 Development of decision models for all case studies.

Types & Formats. M4A, MP3, accompanied by transcripts, images and consent forms.

Origin. Collected by the WP6 members.

Data utility. Due to the privacy issues, only the transcripts will be made publicly available. They can be reused by the other researchers on the topic in the particular locations.

2.7 Questionnaire-based surveys

Purpose. The surveys will mostly serve the following task:

- T 3.3 Surveys of households and companies.

Types & Formats. CSV, DOCX and TXT files.

Origin. Collected by the WP3 members.

Data utility. Anonymised surveys can be reused by other researchers investigating socio-cultural and socio-economic factors that influence metabolic patterns in the case study.

2.8 Socio-economic statistics

Purpose. This type of data will serve the purpose of background data to support the following tasks:

- T 2.2 Testing and Implementing the GDSE in the PULLs;
- T 3.1 Spatial analysis;
- T 3.2 Material flow analysis;
- T 3.3 Surveys of households and companies;
- T 4.3 Generating evaluation models;
- T 4.4 Generating impact models;
- T 5.2 and T 5.4 Carrying out of PULLs;
- T 6.1 Analysis of the decision making landscape in the case study areas;
- T 6.2 Development of decision models for all case studies;
- T 6.3 Implementation of decision (support) models;
- T 6.4 Integration of the decision models and the geo-design decision support environment;
- T 7.2 Analysis of the characteristics of the case study areas for the purpose of knowledge transfer.

Types & Formats. XLSX and CSV files mostly.

Origin. Collected from global, national and regional databases, reports and provided directly by the project partners.

Data utility. Unless generated by combining multiple data sources to provide new insights, data that is already made available through public data repositories will not be preserved for long term in order to optimise resource use and avoid redundancy. This should be the case with most of the socio-economic data used within REPAiR.

2.9 Expected data size

The total estimated data size is less than 100GB per case study, so the overall data size is expected to stay below 1 TB in total. The size of the code base: 0.5 GB (negligible).

3 Data Storage and Backup

3.1 OSF Storage

For OSF Storage, files are stored in multiple locations and on multiple media types. Three types of hashes (MD5, SHA-1, SHA-256) for files are kept. OSF keeps parity archive files to recover from up to 5% bit error. They use Google Cloud[1] for active storage and Amazon Glacier[2] as a backup location. File backups are hosted at Glacier, and there are daily backups on Google Cloud for 60 days. Please refer to Google Cloud and Glacier documentation for details about the other robustness features they provide.

Further, the OSF database is backed up via streaming replication 24 hours a day, and incremental restore points are made twice daily. The OSF database is maintained in encrypted snapshots for an additional 60 days. Database backups are verified monthly. Operational data (e.g., config files) for other OSF services are backed up in primary cloud file storage for 60 days. Logs are primarily stored in Google Cloud cold storage indefinitely. In certain cases a third party aggregation service is used for up to 90 days, then backed up to Amazon S3[3] indefinitely (OSF, 2018[4]).

3.2 TU Delft Server

There are two servers on site of the TU Delft campus that are configured for the REPAiR project. The production server is responsible for making the GDSE web available, including all the relevant datasets. Additionally, the Geoserver is running on the production server providing WMS and WFS-Layer for the case studies.

The backup server is responsible for storing copies of the data sets and configuration files from the production server. Furthermore the backup server is running a replica of the GDSE-Backend, the databases and the Geoserver found on the production server. These replicas allow to quickly switch to the backup server in case the production server gets damaged (during a PULL workshop, for instance). Currently most of the collected datasets are stored on the OSF platform which implements its own backup procedure. These datasets are, following the project progress, being preprocessed and then migrated in a form of structured database entries to the GDSE (production server). The backup and restore procedure is configured. It creates daily snapshots of the databases and of all configuration files on the production server. The restore scripts have been tested..

3.3 4TU Archive

The 4TU.Centre for Research Data is a certified and trusted repository, that has a stable infrastructure for 10 years now, it holds the Data Seal of Approval. All

[1] http://aws.amazon.com/glacier/
[2] http://aws.amazon.com/glacier/
[3] https://aws.amazon.com/s3/
[4] http://help.osf.io/m/faqs/l/726460-faqs

datasets are mirrored and saved in two additional server spaces in the Netherlands to consolidate secure and safe long-term data storage. After 4TU.Centre for Research Data will make its restricted access feature public, the project members will consider depositing sensitive data.

According to the preservation strategy of 4TU.Research Data the data files will be usable even after a long period of time. To ensure reproducibility and possible interoperability it is crucial that every dataset is accompanied by a detailed documentation in form of the readme.txt file. This documentation provides the necessary insights for future researchers about methodologies, approaches, decision-making and software used. The readme.txt files will be generated automatically at the end of the research by using OSF Wiki descriptions and metadata fields. Using non-proprietary data formats will also reduce the risk of long-term expiry.

4 Data Documentation (Metadata)

All the data used in REPAiR project will be stored in OSF repository. The repository can be used to share manuscripts, unpublished findings, data, and work in progress as it makes all of it citable. The discoverability of the work is ensured through tagging items (i.e. datasets, documents, code, etc.) with relevant keywords, that are automatically indexed by the search. Every project, component and file on the OSF has a persistent and unique URL, that can be made accessible either publicly, with restrictions or kept private. Public projects can be given a DOI (Digital Object Identifier). All DOIs generated on the OSF are free for users.

4TU.Centre for Research Data applies the Dublin Core metadata standard, that covers basic information about the dataset, additionally a readme.txt file with detailed documentation is demanded to provide further information. The DOI-system is integrated to enable persistent and unique identification of the datasets. Using the ORCID-system (Open Researcher and Contributor ID) enables the correct and unambiguous allocation of people related to the specific research outcome.

OSF provides an automated version control, logging by whom and when changes were made, and storing the previous versions. Therefore files with the same type of content should not include version numbers, dates, user initials, words as `final` , `first` , `draft` , etc. Instead they should be named exactly the same as the previous version of the file.

The structure of the OSF storage is as following:

Project → Component → Folder → File

There will be only one project called REPAiR;

There will be 6 components for every case study: Amsterdam, Ghent, Hamburg, Łódź, Naples, Pécs and one component for Common data. Components have their own privacy and sharing settings, contributors, access rights, as well as their own unique, persistent identifiers for citation, and their own wiki and add-ons. Each of the case study components will have subcomponents that correspond to different types of data used and generated in REPAiR as in Section 3 Data Collection and Generation.

Every subcomponent will have folders that correspond to different datasets. The folders may have different subfolders for different parts of the dataset. The folder should follow the task structure of the proposal, their name should therefore start wit TX.X and be followed by the keywords of the task. e.g. `T8.5_data_management`. The subfolders do not need to follow the same naming conventions.

A few other notes on file naming conventions:

- "_" and "-" should be used in file names instead of spaces to deliminate units of metadata;
- "-" is used for words that need to be globbed together, "_" separates different information units;
- no punctuation;
- no special characters (e.g. $, @, %, #, &, *, (,), !);
- a new file with the same name as an existing file will automatically replace the existing file and a downloadable copy of the replaced file will be saved;
- for raw data "_raw" is added at the end of the filename and the file is made read-only;
- specific naming conventions are indicated in component's Wiki.

Each dataset used in the REPAiR project will be listed in the DMP according to the following data description (general metadata) template. The metadata will be stored in a metadata.txt in each folder of a separate dataset.

Field	Description
Dataset	Name of the dataset. Dataset title must always begin with the task code it is associated with and end with the relevant case study and keyflow. e.g "T3.2 Actors in Plastic and Packaging flows in Pecs"
Creator	Might be a person, a group of people or an institution
Description	A short description of the dataset (should include information on how the data was generated, what parts does it consist of, quickly introduce software or experiment setup)
Keywords	Keywords that will help indexing, tagging and finding the data once it is made publicly available.
Data purpose	The purpose of the data collection/generation and its relation to the objectives of the project. The concrete task, milestone or deliverable should also be mentioned if relevant.
Language	English/Dutch/Italian/etc.
Data Collection Period	Indicates the start and end of data collection. If data has been derived from other sources, the period indicates data production rather than data collection.
Data Type	Observational *(captured in real time, typically cannot be reproduced exactly. Examples: sensor readings, survey results, images)* / Experimental *(from labs and equipment, can often be reproduced. Examples: material composition, recyclability)* / Simulation *(from models, can typically be reproduced if the input data is known. Examples: climate models, economic models, biogeochemical models)* / Derived (*after data mining or statistical analysis, can be reproduced if well documented. Examples: compiled database, 3D models*) Designed (*data generated using creative processes, cannot be reproduced. Examples: eco-innovative solutions, software code, reports*)
Data formats	Text / numbers / images / 3D models / code / audio files / video files / reports / surveys / maps / scientific articles .txt; .csv; .shp; .wkt; .wmv; .zolca; .pdf, .stadat, .sav,...

Data origin	Collected by REPAiR/ generated using data collected by REPAiR / reused from other sources (indicate which) / generated reusing data from other sources (indicate which)
Data Source	If data has been generated or reused, the original source(s) of the dataset must be indicated. That can be an email correspondence, paper, report, public dataset, etc. ideally, a DOI, PID or URL can be added to the original source.
Software	If viewing or editing the data requires a specific software, it must be indicated which. If multiple are available, one is sufficient with preference to the Open Source or Free ones.
Data size	Estimated (known) data size
Stability	Fixed (never change after being collected, generated) / Growing (new data may be added, but old data is never changed, deleted) / Revisable (new data may be added, old data may be changed, deleted)
Confidentiality	Public / Internal /Restricted (indicate restrictions, e.g. embargo period) / Confidential
Archiving	Data should be preserved for long-term archiving / It is sufficient to keep the data only until the end of the project.

Table 4.1. *A documentation table that is used to describe the metadata of each distinct dataset within REPAiR project. The table will be included as a separate metadata.txt file next to each dataset.*

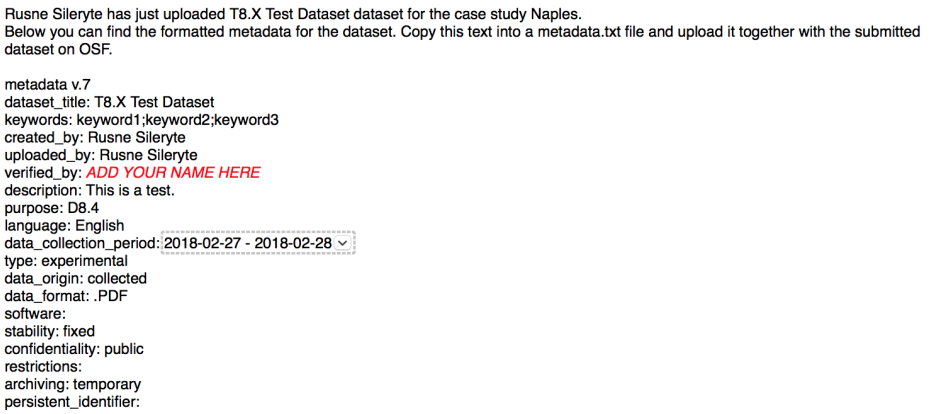
For each dataset that is uploaded to OSF, a metadata form needs to be filled in by the responsible researcher and verified by the Data Captain of the respective case study. It is up to the researcher to decide what constitutes a "single homogeneous dataset" . In some cases it may be a single file, while in others - a collection of multiple files and folders.

A general rule of thumb for the REPAiR project would be that a single dataset:

- spans only one case study (unless it covers all case studies (e.g. European datasets));
- serves only one task or, if it serves multiple tasks, then files cannot be easily separated according to which task they serve;
- has been collected/generated/designed using the same methodology and tools;
- has the same origin, i.e. has been collected, generated or reused from another source;
- if reused from another source, then only a single source has been used, or the same set of sources has been used for all files.

A single dataset does not necessarily need to:

- be collected/generated/designed by a single person;
- have the same file format or single software;
- have the same confidentiality status;
- be archived all together (e.g. in case of an interview dataset, only transcripts may be archived, while video/audio recordings may not).


Rusne Sileryte has just uploaded T8.X Test Dataset dataset for the case study Naples.
Below you can find the formatted metadata for the dataset. Copy this text into a metadata.txt file and upload it together with the submitted dataset on OSF.

metadata v.7
dataset_title: T8.X Test Dataset
keywords: keyword1;keyword2;keyword3
created_by: Rusne Sileryte
uploaded_by: Rusne Sileryte
verified_by: *ADD YOUR NAME HERE*
description: This is a test.
purpose: D8.4
language: English
data_collection_period: 2018-02-27 - 2018-02-28
type: experimental
data_origin: collected
data_format: .PDF
software:
stability: fixed
confidentiality: public
restrictions:
archiving: temporary
persistent_identifier:

Figure 4.1. *An example of the metadata summary email.*

The online metadata form can be filled in using this link:

https://goo.gl/forms/0P823E8mLGbR28sY2

Once a metadata form has been filled in, the Data Captain and the Assistant of the chosen case study will receive an automated email with the metadata summary as in Figure 4.1.

The Data Captain need to copy the email text starting with `metadata v.X` into an empty metadata.txt file and add his/her name instead of the red letters `ADD YOUR NAME HERE`. By doing this he/she confirms that it has been checked and verified that the metadata form follows all the rules set by Data Management Plan.

If the uploaded files have a type-specific metadata file with additional details that were not specified by the general form (e.g. in case of spatial data or interviews), the file(s) need to be uploaded into the same folder as well. Additionally some datasets will have content specific metadata, e.g. geographical datasets. The content specific metadata should adhere to the national standards as well as European INSPIRE Directive.

Long-term, the datasets will be given the metadata used by 4TU.Centre for Research Data, which adheres to the Dublin Core standards. The Dublin Core[1] Metadata Element Set is a vocabulary of fifteen properties for use in resource description.

[1] http://www.dublincore.org/documents/dces/

5 Data Access

There are 4 levels of data confidentiality in the project:

Public - the data can be made public at any time during or after the project;

Internal - the data is shared only between project members, cannot be made public during or after the project;

Restricted - access is restricted due to a period of embargo or if only certain members of the consortium can access it;

Confidential - the data is only accessible by a certain group of people: the respective Data Captain and OSF administrators.

The level of confidentiality is decided by the researcher who is delivering the data to the Data Captain by indicating it in one of the metadata fields.

Although OSF provides a user API for automatically accessing and transferring data from the OSF storage to the TU Delft server, this step will not be fully automated and will need to be supervised by the Data Captains or Data Administrators. This will ensure that access to sensitive data will not be leaked accidentally through automated procedures.

During the project public data will be made accessible through OSF platform. After the project is finished, the data will be made accessible via the web interface of the 4TU.Centre for Research Data [1].

Software code will stay accessible through an open GitHub repository.

Currently manual search, access and download is enabled by the 4TU interface. They are working on an Application Programming Interface (API) to provide advanced services. The access and download does not need a registration or login, since 4TU offers Open Access to the public as well.

The 4TU.Centre for Research Data was already supporting the REPAiR consortium during the proposal stage and is doing so as well during the process, experts from The 4TU.Centre for Research Data were involved in setting up this DMP.

At this moment data access committee is not needed; but the six data captains in the case studies could act as one in the future.

4TU currently has a license that is equal to CC-BY, and they will implement a standardized license soon. The Archive is crawlable, but not the deposited data and

[1] http://researchdata.4tu.nl/home/

`readme` file. That is why basic information and the description are shown in the appropriate metadata fields. 4TU currently offer open access (to the public as well). Future restricted access can be determined by pre-set roles.

Project Leaders at TU Delft as well as GGR and delegated Data Captains are mainly responsible for controlling data access.

At all times there will be more than one dedicated Data Administrator who has full access rights to all the data available in REPAiR. This will make sure that the data will be accessible in case of staff changes, illness etc.

6 Data Sharing and Reuse

4TU.Centre for Research Data enables interoperability and reusability by applying standardized metadata, as well as internationally accepted preferred formats for the deposited data files.

Non-proprietary or open standard file formats will be used to ensure accessibility and reuse.

- For spreadsheets: Comma Separated Values (.csv) or Tab Separated Values (.tsv)
- For text: plain text (.txt), or if formatting is needed, PDF/A (.pdf)
- For presentations: PDF/A (.pdf)
- For images: TIFF (.tif, .tiff), or PNG (.png)
- For videos: MPEG-4 (.mp4)
- For maps: ESRI shapefiles (.shp, .shx, .dbf)

Standard vocabularies will not be implemented in the data management. Providing mappings to more commonly used ontologies is also not intended.

All data that can be open will be released via Creative Commons license CC-BY or CC-0 (according to the H2020 guidelines).

Dataset that have been tagged as `Public` will be made available on OSF platform immediately after upload. If certain restrictions apply, the datasets will be released following the restrictions. Those restrictions will be decided by the responsible researcher delivering the data and verified by the respective Data Captain.

In principle all data will be reuseable by third parties unless special restrictions apply which will be indicated in the accompanying metadata.txt file. It is intended that the data remains reusable for at least 15 years.

Deliverables D9.1 to D9.7 explain any ethical or legal issues that can have an impact on data sharing. The informed consent for data sharing and long term preservation will be included in questionnaires and interviews dealing with personal data.

7 Data Preservation and Archiving

It would be too expensive to preserve all the material produced by the project for the long term, therefore a careful selection needs to be made of what needs to be preserved and what can be disposed. Within REPAiR this decision is made by the researchers providing the data to the Data Captains. The decision is verified by the Data Captain.

The general rules of thumb for long term data archiving in REPAiR project are as following:

- If the original data is collected by the REPAiR team, it needs to be preserved;
- If the data is generated using data collected by the REPAiR team, it needs to be preserved as well;
- If the data is generated using other data sources, that have been reused within REPAiR, then the decision about archiving needs to be based on the ease of replicating the dataset: if it is possible to completely replicate exactly the same dataset, then only the metadata and data generation process (e.g. in a form of code) is sufficient to preserve.
- If the data is reused from other data sources that do not have a persistent identifier and run a risk to become inaccessible in the future, this data should be preserved as well.
- If the data is reused from other data sources, that have a persistent identifier, then it is sufficient to preserve the metadata of such datasets.
- If datasets have been cited by any of the publications, reports or deliverables, they need to be preserved long term.

The estimated total costs for archiving the data in the selected repository are min. 6.000 Euro.

It is assumed that OSF repository, TU Delft servers and 4TU archive are all sufficiently safe and do not need to be replicated elsewhere.