Waste Recycling Management Analysis

WRMA

Data Management Plan (DMP)

Lead partner:	TU Wien
Version:	1.0
Status:	WIP
Dissemination level:	PU: Public
Document link:	https://github.com/BubuGly/194.044-DaStVO-2024S

Deliverable abstract
This deliverable is the Data Management Plan (DMP) for the Waste Recycling Analysis project that outlines the strategies for managing data collected and generated during the research. The objective of this DMP is to ensure that the data handling practices adhere to the principles of findability, accessibility, interoperability, and reusability (FAIR), thereby maximizing the value and impact of the data.

Waste Recycling Management Analysis applies for funding from the European Union's Horizon Europe research and innovation Programme.

DELIVERY SLIP

	Name	Partner/Activity	Date
From:	Ioan Gulyas	Researcher	30.04.2024
Moderated by:			
Reviewed by:			
Approved by:			

DOCUMENT LOG

Issue	Date	Comment	Author
1.0	30.04.2024	Initial Version	Ioan Gulyas

TERMINOLOGY

https://eosc-portal.eu/glossary

Acronym	Definition
DMP	Data Management Plan
CSV	Comma Separated Values
EOSC	European Open Science Cloud
FAIR	FAIR-Principles: Findable, Accessible, Interoperable, and Reusable
MB	Megabyte
PDF	Portable Document Format
WP	Work Package
XML	Extensible Markup Language
JSON	JavaScript Object Notation
EUROSTAT	Statistical Office of the European Union

Content

1. Executive Summary

1.1. Introduction

This document outlines the Data Management Plan (DMP) for the Waste Recycling Analysis project, which aims to evaluate and optimize waste recycling processes using data-driven insights. This project will involve the collection, processing, and analysis of data related to waste types, recycling methods, and efficiency rates.

1.2. Data Summary

- 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.
- What types and formats of data will the project generate or re-use?
- What is the purpose of the data generation or re-use and its relation to the objectives of the project?
- What is the expected size of the data that you intend to generate or re-use?
- What is the origin/provenance of the data, either generated or re-used?
- To whom might your data be useful ('data utility'), outside your project?

In the Waste Recycling Analysis project, I will exclusively re-use existing data obtained from public databases and other accessible sources (EUROSTAT). This data primarily consists of municipal waste composition and recycling rates, sourced from publicly available repositories. The decision to solely re-use existing data was made based on its relevance to the project's objectives and the availability of comprehensive datasets.

The types of data to be re-used include detailed records of waste composition, recycling methods, and efficiency metrics from various municipalities. These datasets are crucial for benchmarking current recycling practices and identifying areas for improvement.

The formats of the re-used data include structured datasets, typically in CSV or similar formats, ensuring compatibility with common data analysis tools and facilitating integration into my analysis workflow.

The purpose of re-using this data is to provide a robust context for my analysis and to leverage existing knowledge in the field of waste management. By re-using publicly available data, I aim to contribute to the broader understanding of recycling processes and to identify best practices for sustainable waste management.

The expected size of the re-used data is relatively small, as it encompasses comprehensive records from multiple municipalities over extended periods, but at a small storage cost. While exact estimates vary, I anticipate working with datasets totaling several hundreds of MB in size.

The origin/provenance of the re-used data is well-established, originating from publicly accessible repositories and databases maintained by governmental and research

institutions. This ensures the reliability and integrity of the data, critical for my analysis and research outcomes.

Outside of my project, this re-used data holds utility for a wide range of stakeholders, including environmental researchers, policymakers, and waste management professionals. By making this data accessible and contributing to its analysis, I aim to support evidence-based decision-making and promote sustainable practices in waste management on a broader scale.

Produced datasets:

dataset ID	title	type	format	estimated volume	contains sensitive data
P1	Eurostat	Eurostat_final.csv	CSV	-	No
P2	CO2_Greenhouse	co2_emission.csv	CSV	-	No
P3	Waste Management	Waste_management_final.csv	CSV	-	No

1.3. Methods and software used for data generation and reuse

In the Waste Recycling Analysis project, multiple techniques will be used. Given the project's focus on analysis, data will be collected, cleaned and brought to a common source in order to have easy access for performing the analysis on it.

Python will serve as the primary programming language throughout the project. Python's versatility and rich ecosystem of libraries for data manipulation and analysis make it well-suited for this task. Being a common choice in the data field, in this case it is also suitable given the multitude of tools that can be used and deployed using Python for statistical analysis, data manipulation and analysis.

For reproducibility and ease of use, Jupyter Notebooks will be the main way of storing the technical part, meaning the code. This allows any open-end user to reuse and refactor the code in an easy way and it also allows me to customize and adapt the code base and the requirements over time as needed and without any hardware limitations occurring.

This combination of methods and tools will enable a comprehensive analysis of the data collected, leading to valuable insights and recommendations for enhancing waste recycling processes.

1.4. Foreseeable research uses and /or users.

Data will benefit researchers and practitioners in environmental science, public policy, and waste management sectors.

2. FAIR data

2.1. Making data findable, including provisions for metadata

- Will data be identified by a persistent identifier?
- 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.
- Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?
- Will metadata be offered in such a way that it can be harvested and indexed?

In the Waste Recycling Analysis project, ensuring the findability of data is paramount to facilitate its discovery and potential re-use. To achieve this:

Persistent identifiers, such as Digital Object Identifiers (DOIs), will be assigned to all generated and reused data. These identifiers will uniquely distinguish each dataset, ensuring their long-term accessibility and retrievability.

Rich metadata will accompany the datasets to provide comprehensive descriptions facilitating discovery. Following standards such as the Dublin Core Metadata Initiative (DCMI), metadata will include essential details such as title, creator(s), creation date, description, keywords, license information, data format, provenance, and related publications. Additionally, domain-specific metadata elements will capture pertinent information specific to waste management and recycling processes. In cases where standardized metadata fields are lacking, custom metadata elements will be defined to ensure essential information is captured.

Keywords will be carefully selected and included in the metadata to optimize discoverability and potential re-use. These keywords will accurately reflect the content and scope of the datasets, aiding users in locating relevant data through search queries.

Metadata will be offered in standardized formats, such as XML or JSON, enabling harvesting and indexing by data repositories and search engines. By encoding metadata in machine-readable formats, we ensure interoperability with various data management systems, enhancing the visibility and accessibility of the data to potential users.

Overall, these measures will enhance the findability of data in the Waste Recycling Analysis project, facilitating its discovery, access, and potential re-use by researchers, policymakers, and practitioners in the field of waste management and environmental sustainability.

2.2. Making data accessible

Repository:

- Will the data be deposited in a trusted repository?
- Have you explored appropriate arrangements with the identified repository where your data will be deposited?
- Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Data:

- 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.
- 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.
- Will the data be accessible through a free and standardized access protocol?
- If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?
- How will the identity of the person accessing the data be ascertained?
- Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Metadata:

- 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?
- How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?
- 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)?

The data generated and used in the Waste Recycling Analysis project will be deposited in a trusted repository that specializes in hosting research data. The selected repository will have a proven track record of reliability, security, and long-term preservation of data. We will make our data accessible by providing open access to data, wherever possible under the GNU General Public License (GPL) version 3.0. In cases where open access is not feasible, we will provide meaningful metadata along with contact information for access requests. This approach ensures transparency and promotes data sharing while adhering to licensing requirements for data distribution.

dataset ID	access conditions	restrictions / embargo reasons	estimated publication date	location for publication (repository)	PID	license
P1	Open		2024-04-30	Github	Not yet	GPL-3.0

Repository description:

Methods or software needed to access and use data: Potential users may require various software tools to access and reuse the data, as the datasets consist of CSV file formats. While specific software may be needed to open certain file types, Python offers versatility in reading and processing diverse data formats. With Python, users can leverage libraries such as pandas for CSV files, SciPy for MAT files, and PIL or OpenCV for image files like JPG. This flexibility ensures that users can access and manipulate the data regardless of its format, facilitating its reuse for future research purposes, including the final predictive maintenance model and charts generated from the analysis.

Additionally, the entire project, including the datasets, code, and documentation, will be hosted on GitHub. GitHub provides a collaborative platform for version control, issue tracking, and project management. Users can access the project repository on GitHub, clone it locally, and utilize the provided scripts and notebooks to analyze the data, develop machine learning models, and reproduce the results. This centralized platform enhances accessibility and transparency, enabling seamless collaboration and contribution from the research community.

2.3. Making data interoperable

- 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?
- 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?
- Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

Dublin Core Metadata Initiative (DCMI): Utilizing Dublin Core metadata elements to describe the essential attributes of our datasets, ensuring compatibility and consistency in metadata representation.

Data Documentation Initiative (DDI): Adhering to the DDI standard for describing data produced from social, behavioral, and economic research, ensuring comprehensive documentation of data variables, collection methods, and processing procedures.

International Standardization Organization (ISO) Standards: Following ISO standards such as ISO 19115 for geospatial metadata and ISO 2709 for bibliographic information exchange, ensuring compatibility with international metadata standards.

Resource Description Framework (RDF) and Linked Data Principles: Embracing RDF and linked data principles to represent data and metadata in a machine-readable format, enabling semantic interoperability and integration with external datasets and knowledge graphs.

2.4. Increase data re-use

- 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.)?
- 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?
- Will the data produced in the project be useable by third parties, in particular after the end of the project?
- Will the provenance of the data be thoroughly documented using the appropriate standards?
- Describe all relevant data quality assurance processes.
- Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

Comprehensive documentation will be provided to validate data analysis and facilitate data reuse. This documentation will include readme files containing information on the methodology, codebooks, data cleaning procedures, analytical techniques employed, variable definitions, units of measurement, and any other relevant details necessary for understanding and replicating the analyses conducted.

The data produced in the project will be made freely available in the public domain to permit the widest possible reuse. These datasets will be licensed using standard reuse licenses, such as Creative Commons licenses, in line with the obligations set out in the Grant Agreement. By adopting open licensing frameworks, we aim to maximize the accessibility and reuse potential of the data while ensuring compliance with legal and ethical requirements.

The data produced in the project will be designed to be usable by third parties, particularly after the end of the project. Clear documentation, standardized formats, and open access

policies will facilitate the usability of the data by diverse stakeholders, including researchers, policymakers, industry practitioners, and the general public.

The provenance of the data will be thoroughly documented using appropriate standards and best practices. This documentation will include information about the origin of the data, data collection methods, data processing steps, and any transformations applied to the data. By transparently documenting data provenance, users can assess the reliability and trustworthiness of the data for their intended purposes.

Relevant data quality assurance processes will be implemented to ensure the integrity and reliability of the data. These processes will include data validation checks, consistency checks, outlier detection, and error correction procedures. Additionally, data quality metrics will be defined to assess the completeness, accuracy, and consistency of the data. Regular audits and reviews will be conducted to maintain data quality throughout the project lifecycle.

In addition to addressing the FAIR principles, the Data Management Plan (DMP) for the Waste Recycling Analysis project will carefully consider aspects related to the allocation of resources, data security, and ethical considerations. By adopting a holistic approach to data management, we aim to maximize the value, accessibility, and ethical integrity of our research outputs.

3. Other research outputs

- In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).
- Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

Identifier Assignment: Each digital and physical research output will be assigned a persistent identifier, such as a DOI or a unique accession number, to ensure its traceability and citability.

Metadata Provision: Comprehensive metadata will be created for each research output, including detailed descriptions, relevant keywords, creator information, and licensing terms. These metadata will be made openly available to enhance the discoverability and reusability of the outputs.

Accessibility: Efforts will be made to ensure that research outputs are accessible through free and standardized access protocols. Digital outputs, such as software, workflows, and

models, will be made available through online repositories or platforms, while physical materials will be stored in accessible repositories or distributed through established channels.

Long-term Preservation: Long-term preservation strategies will be implemented for both digital and physical research outputs to ensure their continued availability and usability beyond the project's duration. This includes regular backup procedures, archival storage solutions, and periodic review of preservation plans to adapt to evolving technologies and standards.

4. Allocation of resources

- 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.)?
- 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)
- Who will be responsible for data management in your project?
- 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)?

In the Waste Recycling Analysis project, the costs associated with making data and other research outputs FAIR encompass both direct and indirect expenses. Direct costs include expenditures related to data storage, archiving, metadata creation, and security measures. These costs will vary based on the volume of data generated, the complexity of metadata requirements, and the chosen storage and security solutions. Indirect costs may encompass personnel expenses for data management activities, training, and compliance monitoring. Overall, the estimated costs for ensuring data FAIRness in the project are expected to be moderate, reflecting the project's commitment to open science and data sharing.

Covering these costs will be facilitated by the Horizon Europe grant allocated to the Waste Recycling Analysis project. As per the Grant Agreement conditions, expenses related to research data and output management, including efforts to make data FAIR, are eligible for funding. The project budget includes provisions specifically allocated for data management activities, ensuring adequate resources are available to effectively implement FAIR principles.

Responsibility for data management in the project rests with a dedicated data management team comprising project members with expertise in data science, information management, and domain-specific knowledge in waste recycling. This team will oversee all aspects of data management, including data collection, storage, documentation, sharing, and preservation, ensuring compliance with FAIR principles and project objectives.

Long-term preservation of research data will be ensured through the implementation of robust data archiving and preservation strategies. The necessary resources for this endeavor, including costs and potential value, will be determined by the data management team in collaboration with project stakeholders. Decisions regarding what data to keep and for how long will be guided by project requirements, data relevance, and legal obligations. The data management team will ensure that data preservation activities align with FAIR principles and contribute to the long-term accessibility and usability of research outputs beyond the project's duration.

5. Data security

- 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)?
- Will the data be safely stored in trusted repositories for long term preservation and curation?

5.1. Storage and backup facilities

Storage and backup of data will be performed locally and on GitHub. As a measure of extra caution, the data will also be store on the platform of the Technical University of Vienna (TUWEL).

5.2. Data security and protection of sensitive data

Adherence to GDPR, with anonymization of personal data where applicable. Access to data during research:

dataset ID	selected project members	all other project members	the public
D1	writing	writing	reading only

All incidents will be handled individually and in person.

Long-term preservation and deletion of data

dataset ID	location for long-term storage	minimum retention period (≥ 10 years)
D1	Github	10 years

6. Ethical and legal issues

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

6.1. Personal data

No personal data will be involved in the development of the project.

6.2. Intellectual property rights and ownership

There are not conflicts in regard to intellectual property rights and ownership. All the data used is publicly available and free to use.

6.3. Ethical issues

No ethical issues will arise from the project at hand.

7. Other issues

- Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?