r3volution

reuse · resource · recovery

Deliverable 7.5.
Data Management Plan
v1

r3volution.eu

Funded by the European Union

European Commission

Data Management Plan v1

D7.5: Data Management Plan v1

Summary

The present document defines the R3VOLUTION Data Management Plan (DMP): it describes the data management life cycle for the data to be collected, processed, and/or generated during the project execution, including data management strategy, naming and metadata guidelines. It defines best practices and regulatory requirements, and facilitates efficient data management so as to enable effective collaboration and compliance with said best practices and requirements.

This document corresponds to the first version of the Data Management Plan that will be continuously reviewed and updated throughout the project lifetime as it accrues data, at each reporting period (M18, M36, M48) and will be resubmitted as D7.6, D7.7, D7.8

Deliverable number	Work Package
D7.5	WP7
Lead beneficiary	**Deliverable author(s)**
CETAQUA	Clothilde Breger
Quality assurance / reviewers	
CERTH CETAQUA	Grigorios Gtzionis Eric Santos Clotas
Planned delivery date	**Actual delivery date**
30/06/2024	27/06/2024
Dissemination level	x PU = Public □ SEN = Sensitive, only for members of the consortium

Table of Contents

List of Tables

Executive summary

The present document defines the R3VOLUTION Data Management Plan (DMP): it describes the data management life cycle for the data to be collected, processed, and/or generated during the project execution, including data management strategy, naming and metadata guidelines. It defines best practices and regulatory requirements, and facilitates efficient data management so as to enable effective collaboration and compliance with said best practices and requirements.
This document corresponds to the first version of the Data Management Plan that will be continuously reviewed and updated throughout the project lifetime as it accrues data, at each reporting period (M18, M36, M48) and will be resubmitted as D7.6, D7.7, D7.8

List of Acronyms and Abbreviations

AI	Artificial Intelligence
CV	Computer Vision
DMP	Data Management Plan
DT	Digital Twin
EC	European Commission
EU	European Union
GA	Grant Agreement
HE	Horizon Europe
IPP	Intellectual Property Protection
ML	Machine Learning
PC	Project Coordinator
PO	Project Officer
XAI	eXplainable Artificial Intelligence
WWTP	WasteWater Treatment Plant
WP	Work Package
WPL	Work Package Leader

1. Introduction

This document is developed as part of the R3VOLUTION (A rEVOLUTIONary approach for maximising process water REuse and REsource REcovery through a smart, circular and integrated solution) project, which has received funding from the European Union's program HORIZON-CL4-2023-TWIN-TRANSITION-01-40, under the Grant Agreement Number 101138245. The Data Management Plan corresponds to Deliverable 7.5 of Work Package 7 (WP7 - Project management). It describes the data management life cycle for all data sets that will be collected, processed, and generated by the project, and is intended to be read by any member of the consori. Additionally, this DMP indicates how the consortium will handle the research data during and after completion of the project. This is of particular relevance as a large amount of data will be generated from the case studies of this project.

The main objective of the digital tools developed within work package (WP3) is to take advantage of the data available to build ML models for process control and optimisation. Data management within WP3 will collect the necessary data relevant for the baseline definition phase as well as during operation from external public and private databases, sensors and IoT devices (existing and new ones) and specific industrial data management systems in real-time or scheduled interval.
A database management platform will be designed and built by Z-Prime (corresponding to task 3.1) to develop a digital twin with recommender systems of diverse categories such as Explainable AI (XAI), Computer Vision (CV), and virtual sensing. Another platform will be set up and owned by CETAQUA, and both platforms will be accessible to all partners.

This document will first describe a general code of conduct for data management and use (section 2), before listing datasets to be uploaded into the Data Management Platform (section 3). Then section 4 enumerates data policies, that is to say general guidelines on data management, dividing and defining responsibilities. Compliance with FAIR data guidelines is subsequently outlined (section 5), including the definition of the second platform and actions relative to publications. Finally, allocation of resources are briefly commented on (section 6).

The DMP is a dynamic and continuous document that will be updated throughout the lifetime of the project. The revised versions of this document will be published as Deliverable 7.6 (Data Management Plan v2, M18), Deliverable 7.7 (Data Management Plan v3, M36) and Deliverable 7.8 (Data Management Plan v4, M48). In month 8, the Data Management Platform demonstration will be delivered (deliverable 3.1), it will include technical details on how compliance of the ideas described in this document will be achieved.

2. Code of Conduct

One of the most crucial aspects of data management is establishing clear guidelines on permissible data usage from the project's data pool. This can be achieved by defining specific purposes, establishing ethical principles, or creating a general Code of Conduct. The R3volution Code of Conduct was developed based on the European Code of Conduct for Research Integrity (ALLEA, 2017), the International Sociological Association's Code of Ethics (ISA, 2021), and the Data Science Code of Professional Conduct (Data Science Association, 2021). It is tailored to the entire project, covering all data sources and cases across all living labs, with a particular focus on data science approaches. The following sections outline the terminology and rules upon which the R3volution Code of Conduct is based

2.1 Terminology

- **Agency Problem:** A situation where one party (the agent) is supposed to act in the best interests of another party (the principal), but their interests and information levels differ, so the principal cannot be sure the agent is truly acting in their best interests. This can also lead to moral hazard, where the agent lacks incentive to guard against risks they are protected from.

- **Algorithm:** A set of rules or steps to follow to solve a problem or achieve a desired outcome, especially mathematical procedures to calculate a result.

- **Causation:** The scientifically proven relationship between a cause and its effect, where one event leads to or brings about another event as a consequence.

- **Cherry picking:** Highlighting only the individual cases or data points that support a particular position, while ignoring a significant portion of related cases/data that contradicts that position. This can constitute scientific fraud, evidence suppression, or the fallacy of using incomplete evidence.

- **Confidential information:** Any information about project participants or organisations that is not public knowledge and should not be made public without their permission.

- **Correlation:** A statistical relationship between two variables where a change in one is associated with a change in the other.

- **Data:** Raw factual or non-factual information, measurements or statistics recorded in a tangible or digital format that must be processed and analysed to become meaningful.

- **Data mining**: Using advanced data analysis techniques and algorithms to discover patterns and extract new insights from large data sets.

- **Data quality:** Evaluating the accuracy, completeness, and reliability of data.

- **Data Science:** The scientific study of creating, validating, and transforming data to generate knowledge and insights.

- **Fraud:** Deceptive conduct involving falsifying results, fabricating or manipulating data/materials, plagiarising others' work, or violating applicable laws and regulations.

- **Known/Knowingly:** Having actual knowledge about a particular fact, which can be inferred from the circumstances.

- **Knowledge:** Meaningful information backed by scientific evidence.
- **Machine Learning:** A field focused on enabling computers to learn and improve from data without being explicitly programmed.
- **Protagoras problem:** Making distorted assumptions and calling them "science" and "evidence". Sincerity in assumptions is key.
- **Reasonable:** For a researcher's conduct, what a reasonably prudent and competent researcher would do.
- **Spurious correlation:** A correlation between two variables that erroneously appears as a direct (causal) relationship, when it actually arises from their relationships with other variables.
- **Statistically significant:** A statistical measure of whether observations represent an actual pattern rather than just chance, though significance may not imply substantive importance.
- **Statistics:** The field of mathematics involving the collection, analysis, interpretation, presentation, and organisation of data, its ultimate goal being to extract meaningful insights and inform decision-making processes through quantitative data analysis.
- **Substantial:** Of material importance and clear, weighty significance when referring to degree or extent.
- **Variable:** A value that can change within the context of a given problem or set of operations, either independently or dependently.

2.2 Rules

The research within the project should adhere to the principles and practices outlined in the European Code of Conduct for Research Integrity, founded on the core tenets of reliability, honesty, respect, and accountability. Concerning data practices and management, all researchers, institutions, and organisations involved shall:

1. Ensure proper stewardship, curation, and secure preservation of all data and research materials, including unpublished ones, for an appropriate duration.
2. Facilitate open access to data in accordance with the FAIR Principles (Findable, Accessible, Interoperable, and Reusable) whenever feasible, while respecting necessary restrictions.
3. Maintain transparency regarding the means to access and utilise their data and research materials.
4. Acknowledge data as a legitimate and citable research output.
5. Ensure fair provisions for management, ownership and intellectual property of research outputs in any relevant contracts or agreements.

For data science approaches within the project, the following additional guidelines (adapted from the Data Science Association's Code of Professional Conduct) apply:

6. Competent data science research necessitates the knowledge, skill, thoroughness, and preparation reasonably required for the tasks at hand.

7. Researchers shall not engage in or facilitate the collection, processing, or transfer of data known to be associated with criminal or fraudulent activities.

8. Research results derived from data sets must be expressed to an extent reasonably necessary to provide others with a transparent and comprehensible knowledge base for informed decision-making.

9. All information created, developed, received, used, or learned during the project, which is not generally known to the public, shall be treated as confidential. Researchers have a duty to safeguard all confidential information, regardless of its form or format, from its creation or receipt until its authorised disposal. Protecting this valuable asset is critical not only for compliance with legal and ethical requirements, but also to build trust.

10. Researchers may disclose information to the extent reasonably believed necessary to prevent project partners from committing a crime or fraud that is plausibly certain to result in consequential injury to the financial interests or property of another.

11. Researchers must take adequate measures to prevent the inadvertent or unauthorised disclosure of, or access to, information, including but not limited to:

    a. Refraining from displaying, reviewing, or discussing confidential information in public places, in the presence of third parties, or in situations where it may be overheard.

    b. Avoiding the transmission of confidential information outside the organisation or professional practice to personal email accounts or the removal of confidential information by copying it to any form of recordable digital media device.

    c. Refraining from communicating confidential information to individuals not necessarily involved in the task.

12. Researchers shall comply with the procedures and rules defined in the Grant Agreement pertaining to the acceptance, proper use, and handling of confidential information, as well as any additional written agreements.

13. Researchers shall continue to protect confidential information even after termination of work for the partner that provided the information.

14. Upon termination of the project, researchers shall return any and all confidential information in their possession or control.

15. Data sets shall be rated with regard to data quality, and this rating shall be disclosed to enable informed decision-making, as poor or uncertain data quality may communicate a false reality or promote an illusion of understanding. Researchers shall take reasonable measures to protect data users from relying on and making decisions based on poor or uncertain data quality.

16. Similarly, data sets shall be rated with regard to the quality of evidence, and this rating shall be disclosed to enable informed decision-making. Researchers shall take reasonable

measures to protect data users from relying on and making decisions based on weak or uncertain evidence, as evidence may vary in strength or certainty.

17. If a researcher believes a project partner is misusing research results to communicate a false reality or promote an illusion of understanding, the researcher shall take reasonable remedial measures, including disclosure to the project partner. The researcher shall take reasonable measures to encourage the appropriate use of data science by the project partner, including, if necessary, disclosure to the project leader.

18. If a researcher becomes aware that a project partner intends to engage, is engaging, or has engaged in criminal or fraudulent conduct related to the provided research, the researcher shall take reasonable remedial measures, including disclosure to the project leader if needed.

19. A researcher shall not knowingly:

    a. Fail to employ scientific methods in performing data science.

    b. Fail to rank the quality of evidence or the quality of data in a reasonable and understandable manner.

    c. Claim that weak or uncertain evidence is solid evidence.

    d. Misuse weak or uncertain evidence to communicate a false reality or promote an illusion of understanding.

    e. Fail to rank the quality of data in a sound and comprehensive manner.

    f. Claim that poor or uncertain data quality is good data quality.

    g. Misuse poor or unsure data quality or data science results to communicate a false reality or promote an illusion of understanding.

    h. Fail to disclose any and all data science results or present incomplete evidence as real data science evidence. A researcher may present a theory constituting incomplete evidence so long as it is labelled and clearly communicated as such.

    i. Cherry-pick data and data science evidence. Doing so could constitute scientific fraud, suppressing evidence, or the fallacy of incomplete evidence.

    j. Fail to attempt to replicate data science results.

    k. Fail to disclose unsuccessful replication of data science results, failed experiments or disconfirming evidence known to the researcher to be directly adverse to the researcher's position.

    l. Offer evidence that the researcher knows to be false. The researcher must disclose any doubts about the quality of data or evidence. If a researcher has offered material evidence and subsequently becomes aware of its inadequacy, they shall take reasonable remedial measures including disclosure to the recipients of the research results. A researcher may disclose and label evidence the researcher reasonably believes is false.

20. Researchers shall exercise reasonable diligence when designing, creating, and implementing algorithms and machine learning systems to avoid harm. They shall disclose any real,

perceived, or hidden risks associated with using the algorithm or ML system. After full disclosure, the user is responsible for deciding whether to use or not use the algorithm. If a researcher reasonably believes an algorithm will cause harm, the researcher shall take reasonable remedial measures and measures encouraging appropriate use of the algorithm or ML system.

21. A researcher shall use their reasonable best efforts to assign value and meaning to the following concepts when conducting data science and communicating results, within or outside of the consortium:

    a. "Statistically Significant"

    b. "Correlation"

    c. "Spurious Correlation"

    d. "Causation"

22. Researchers shall make reasonable efforts to question assumptions and avoid distorting assumptions into “science” and “evidence” (also known as the “Protagoras Problem”).

23. Researchers shall recognize “agency problems” when conducting data science and disclose all conflicts of interest by informing the project coordinator and the affected project partners to develop a mutually acceptable solution.

24. Researchers shall exercise reasonable diligence to detect, recognise, disclose, and account for real, perceived, and potentially hidden risks in using data science. They should understand that data creators and system designers may have more knowledge and could hide risks within the data. Researchers shall take reasonable remedial measures, including disclosing these risks to the recipients of their research results.

25. Researchers shall adhere to the data science method, which consists of the following steps:

    a. Careful observation of data, datasets, and relationships between data.

    b. Deduction of meaning from the data their relationships.

    c. Formation of hypotheses.

    d. Experimental or observational testing of the validity of the hypotheses. To be termed scientific, a method of inquiry must be based on empirical and measurable evidence subject to specific principles of reasoning.

3. Data sets to be gathered and processed

3.1 Introduction

Most of the data used within the R3volution project will be uploaded into the Data Management platform resulting from task 3.1. This system will collect, store, and homogenise raw and processed data in various formats from external public and private databases, sensors, and IoT devices from different manufacturers, as well as specific industrial data management systems, in real-time or at scheduled intervals. Designed to be scalable and expandable, the system will accommodate increasing data volumes, new functionalities, and additional services. Appropriate security protocols, such as HTTPS and SSL, along with firewalls, will be implemented. To ensure interoperability, data integration, and efficient querying within the data ecosystem, a generic metadata model will be developed and implemented.

This section enumerates the datasets to be gathered in the Data Management Platform. Their attributes are summarised in a table, including how the data will be collected and for which task it will be used. Thus, the following subsection describes possible use cases of a dataset from each industrial pilot so as to be able to refer to them in a more simple manner in the final sub-section, the table summarising datasets in the platform.
By nature, the data from industrial pilots will mostly be confidential and reserved for use within the project, unless otherwise specified. However, metadata will be collected and stored in another database platform managed by Cetaqua, as explained in section 5 of this document.

3.2 Use cases for each pilot

The following section lists the possible use cases for each industrial pilot's datasets, in alphabetical order Bloom, Celsa, Felix Schoeller and Repsol.

3.2.i Bloom

The removal of organic compounds from the side stream of lignin production can be effectively achieved through the application of combined separation technologies, including Nanofiltration (NF), Reverse Osmosis (RO), and Membrane Distillation (MD). The use cases are defined based on the specific technology employed at each stage, applicable at both laboratory and pilot levels:

- Lab level
    - NF (Nanofiltration) ⇒ use case B1a
    - RO (Reverse Osmosis) ⇒ use case B1b
    - MD (Membrane Distillation) ⇒ use case B1c
- Pilot level
    - NF ⇒ use case B2a
    - RO ⇒ use case B2b
    - MD ⇒ use case B2c

3.2.ii Celsa

To develop a digital twin of an ultrafiltration (UF) and NF pilot plant for the recovery of water from a wastewater source at CELSA (use case C1). At the time of writing this document, Celsa and the UPC are still defining use cases in more detail; this site being a special situation as these partners need to externalise the creation of their pilot.

3.2.iii Felix Schoeller

Use case F1: Explanation of predictive control of WWTP chemical dosing

- Modelling dynamic behaviour of parameters inside the wastewater treatment plant (WWTP), reacting on parameters in 2 different main feed flows.
- Collection of data, proposal for automatic dosing

Use case F2: Thermal simulation of WWTP discharge water temperature

This will be done based on four main impact parameters

- 2 different inlet flows and temperatures
- outside weather conditions
- operation of mechanical and biological treatments
- operation of cooling tower

3.2.iv Repsol

Use case R1: Virtual sensing (T3.4)

This is the practice of replacing a sensor with a Machine Learning algorithm, inferring the amount of component y from measurements of components X1, X2, etc. Results will lead to defining early detection of emerging organic contaminants (y), where a physical sensor would be too expensive or might not even exist at a higher TRL.

Virtual sensors will be developed (trained, validated and tested) by combining historical datasets, real-time data from sensors (yet to be installed in the refinery) and lab measurements of the target organic contaminants.

Use case R2: Computer vision (T3.4)

This will enable early detection of foam episodes (and possibly hydrocarbons in the water). The images will come from 1-2 cameras yet to be installed in the refinery.

Use case R3: Pilot digital twin (T4.2)

Each pilot will have a digital twin in the data platform, to be used for further applications (such as testing treatments) to improve operation of the plant.

3.3 Dataset list

The following table (table 1) lists the datasets that will be collected for each pilot and use case as well as their source, format and method of collection. It will be filled in in more detail as the project advances in future iterations of this document.

Table 1. List of datasets to be collected for the Data Management Platform

No.	Pilot	Use Case	Name of Dataset	Description and purpose	Source of Data	Dissemination level / Licence	Exploitation / Target stakeholders	Embargo Period	DOI
1	Bloom	B1a - Lab NF	Experimental data NF lab	Development of NF process for Bloom upstream case	File	Confidential, internal use only	TBD	N/A	N/A
2	Bloom	B1b - Lab RO	Experimental data RO lab	Development of RO process for Bloom upstream case	File	Confidential, internal use only	TBD	N/A	N/A
3	Bloom	B1c - Lab MD	Experimental data MD lab	Development of MD process for Bloom upstream case	File	Confidential, internal use only	TBD	N/A	N/A
4	Bloom	B2a - Pilot NF	Experimental data NF pilot	Development of NF process for Bloom upstream case	File	Confidential, internal use only	TBD	N/A	N/A
5	Bloom	B2b - Pilot RO	Experimental data RO pilot	development of RO process for Bloom upstream case	File	Confidential, internal use only	TBD	N/A	N/A
6	Bloom	B2c - Pilot MD	Experimental data MD pilot	development of MD process for Bloom upstream case	File	Confidential, internal use only	TBD	N/A	N/A
7	Celsa	C1 - Celsa pilot	Celsa pilot data	Data on UF and NF performance parameters		Confidential, internal use only	Members of the consortium	N/A	N/A
8	Felix Schoeller	F1 - Predictive controlling	Sensor data	Modelling dynamic behaviour of parameters inside WWTP	SQL database, 1 year per file (single input)	Confidential, internal use only	Members of the consortium	N/A	N/A
9	Felix Schoeller	F2 - thermal simulation	Sensor data	Thermal simulation of WWTP discharge water temperature	SQL database, 1 year per file (single input)	Confidential, internal use only	Members of the consortium	N/A	N/A
10	Repsol	R1 - Virtual sensing	Información R3VOLUTION Repsol Puertollano	Human-made measurements of many chemical compounds at relevant points of	API to database	Confidential, internal use only	Members of the consortium	N/A	N/A

No.	Pilot	Use Case	Name of Dataset	Description and purpose	Source of Data	Dissemination level / Licence	Exploitation / Target stakeholders	Embargo Period	DOI
				the refinery water cycle (2021 to 02/2024)					
11	Repsol	R1 - Virtual sensing	Sensor data	Sensor measurements (relevant points for virtual sensing). Most likely two to three different sensors-	API to database	Confidential, internal use only	Members of the consortium	N/A	N/A
12	Repsol	R2 - Computer Vision	Images	Pictures of water in the cycle (relevant points to detect foam issues)	API to Repsol's image database	Confidential, internal use only	Members of the consortium	N/A	N/A
13	Repsol	R3 - Pilot Digital Twin	Pilot data	Data from pilot plant: sensor measurements (different points), activity of different components, ...	API?	Confidential, internal use only	Members of the consortium	N/A	N/A

4. Data policies

The following section outlines the comprehensive framework governing data within the project. It details the governance structure responsible for data access, storage, and security, measures to ensure data integrity, and the integration of a robust Data Quality framework, and defines what the responsibilities for each of these terms entail. This framework ensures the consistent and reliable handling of data across all operational processes.

4.1 Governance structure

Z-Prime being the leader for task 3.1 (creating the Data Management Platform), they are responsible for data integrity and access for all users. As Z-Prime designs and builds the platform, newer versions of this document will be more concrete as to the technical implementation of these requirements.

Each industrial member and their respective innovation partner (Bloom - Vito, Celsa - UPC, Felix Schoeller - Franhaufer, Repsol - CETAQUA) have several roles. Firstly, the industrial companies own any data coming directly from their systems, and as such are expected to provide comprehensive information to their innovation partner about said data, especially relevant to data quality. They should also define who can access their data, along with the innovation partner. Secondly, the innovation partner manages their industrial partner's data, which means that they are the reference point for data integration and documentation for the rest of the consortium. This includes preparing the metadata to be published by CETAQUA (see section 5). Finally, both the industrial and innovation partners will be users of the data and resulting AI applications, for which they will need to be granted access to the platform.

4.2 Usage

The data in the Data Management Platform will be used to develop ML solutions assisting the industry partner's water treatment cycle, including digital twins for the pilots with recommender systems, explainable AI (XAI) systems, process models, virtual sensing and Computer Vision (CV). Section 3.2 specifies which dataset will be used for which purpose, and will be detailed in future versions of this document.
Section 5 sketches out guidelines on publishing results.

4.3 Access

Z-Prime being responsible for data access means that it must implement comprehensive access management policies to verify the identities of users and grant appropriate permissions, thereby maintaining data confidentiality and integrity. By default, data should be accessible to all members of the consortium, but making at least metadata openly accessible should be strongly encouraged in the spirit of open access research (see section 5.2).

4.4 Integrity

Data integrity refers to the accuracy, consistency, and reliability of data throughout its lifecycle. This responsibility entails not only safeguarding the data during storage with advanced encryption and regular security audits but also ensuring secure data transmission channels to prevent interception during access by authorised parties. Maintaining data integrity involves implementing checks and

validation procedures to detect and correct errors, ensuring the data remains trustworthy and fit for its intended use.

4.5 Integration (Data Quality framework)

Integration, in the context of data management, refers to the process of combining data from different sources and providing users with a unified view of this data. This involves ensuring that data from various systems, databases, and applications can work together seamlessly. Effective integration requires establishing common standards and protocols, harmonising data formats, and ensuring consistent data quality and accuracy. The goal is to enable efficient data sharing, interoperability, and comprehensive analysis.
As data managers each “innovation partner” should specify how and when to integrate different sources, whilst Z-Prime is responsible for enacting any data integration from a technical standpoint.

5. R3volution measures to ensure FAIR data management

FAIR data management refers to a set of principles aimed at ensuring that data is Findable, Accessible, Interoperable, and Reusable. This section defines each term and details the measures implemented to satisfy the corresponding requirements.

5.1 Findable

Data is findable when it is easy to locate for both humans and computers. This is achieved through metadata and unique identifiers. For this project, the data used for machine Learning applications will be stored in the Data Management Platform created by Z-Prime. Then a database will be set up and owned by CETAQUA for this data and more, accessible to all the partners. By using Zenodo as the general-purpose open-access repository, all data will be findable. The data will be managed as defined in table 2 below:

Table 2. List of data and metadata to be collected in the open-access platform

Types of data/research output	Accessibility	Interoperability	Reusability
Data from lab experiments (WP2) and sites operation (WP4)	Members of the consortium through Zenodo	Data from lab experiments (WP2) and sites operation (WP4)	Relevant rights for reuse will be appropriately identified for the data type.
Raw laboratory data		Laboratory notebooks (including e-notebooks) – lab notebook policies are in place within each partner organisation requiring meticulous lab book maintenance and regular sign off	
Qualitative data		eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml)	
Quantitative tabular data with minimal metadata		comma-separated values (CSV) file (.csv) or tab delimited file (.tab)	
Quantitative tabular data with extensive metadata		SPSS portable (.por) file format	
Digital image data		TIFF version 6 uncompressed (.tif)	

Digital video data documentation		MPEG-4 (.mp4)	
Documentation		Portable Document Format (PDF) (.pdf)	

5.2 Accessible

Once found, data must be accessible under clearly defined conditions and protocols. As explained in section 4.3, Z-Prime is responsible for implementing the access protocols for the Data Management Platform. Additionally, the platform containing the metadata described above in table 2 will be accessible to all consortium members.

If a beneficiary intends to disseminate results, they must provide 15 days' notice to other beneficiaries, who may object if they can demonstrate a legitimate interest in protecting their results or background. Partners will secure Intellectual Property Rights (IPR) as necessary to comply with open access requirements; this means asking data owners and managers (the relevant pilot - technical pair). Beneficiaries will select the most appropriate route for open access—either 'green' open access (self-archiving) or 'gold' open access (open access publishing).

R3volution will ensure open access to the digital research data generated and other outputs (e.g., models, algorithms, workflows), with all publications adhering to FAIR principles. At the time of publication, a machine-readable electronic copy of the manuscript will be deposited in a general-purpose open-access repository which complies with EOSC requirements under the European OpenAIRE program operated by CERN. This repository can be created in Zenodo, but should be different from the one described in table 2 above so as to comply with different accessibility requirements (consortium or open-access for publication). This strategy guarantees immediate open access to all peer-reviewed scientific publications produced. Immediate open access to deposited information will be provided under the latest version of the Creative Commons Attribution International Public License (CC BY). All metadata will be covered by a Creative Commons Public Domain Dedication (CC 0) or an equivalent. The Open Science & Research Data Management Task is detailed in T7.5 and will be led by CETAQUA.

5.3 Interoperable

Data should be compatible with other data sets and systems, enabling integration and analysis. As mentioned in section 4.5, data managers (i.e. Cetaqua, Fraunhofer, UPC and Vito) are responsible for clarifying how each dataset can be joined to others both within the data management platform and the Zenodo platform.

5.4 Reusable

To be reused in future research and applications, data should be well-documented and have clear usage licences. As specified in section 4.1 and in table 2 above, each data manager should document the data they are responsible for and give said documentation to Cetaqua to be uploaded to the open access Zenodo database.

6. Allocation of resources – Responsibility & Costs

Z-Prime is task 3.1 leader, i.e. they will build the Data Management Platform. Most other data-related labour falls to CETAQUA, Fraunhofer, UPC and Vito as data managers, as part of their ML or DT tasks. No external costs relative to data management are anticipated.

r3volution

reuse · resource · recovery

Funded by the European Union

European Commission