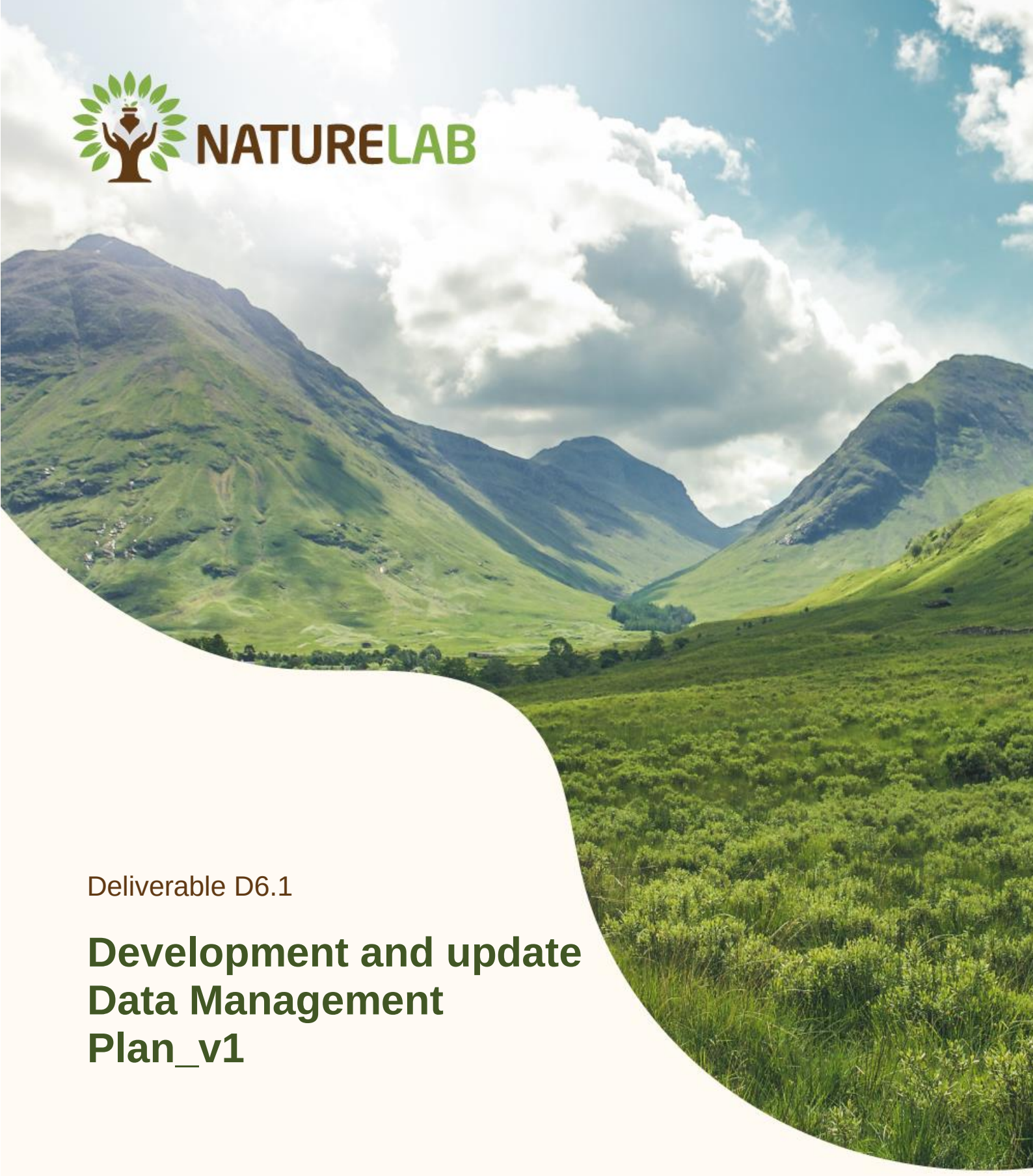

NATURELAB

Deliverable D6.1

Development and update Data Management Plan_v1

Funded by the European Union

UK Research and Innovation

DOCUMENT CONTROL SHEET

PROJECT INFORMATION

Project Number	101083857		
Project Acronym	NATURELAB		
Project Full title	Nature based interventions for improving health and well-being		
Project Start Date	1 June 2023		
Project Duration	54 months		
Funding Instrument	Horizon Europe	Type of action	Research and Innovation Action (RIA)
Topic	HORIZON-CL6-2022-COMMUNITIES-02-02-two-stage		
Coordinator	Laboratório Nacional de Engenharia Civil (LNEC)		

DELIVERABLE INFORMATION

Deliverable No.	D6.1							
Deliverable Title	Development and update Data Management Plan_v1							
Work-Package No.	WP6							
Work-Package Title	Coordination and management							
Lead Beneficiary	LNEC							
Main Author	José Barateiro (LNEC)							
Other Authors	António Antunes (LNEC)							
Due date	M6							
Deliverable Type		Document, Report (R)	X	Data management plan (DMP)		Websites, press & media action (DEC)		Other
Dissemination Level	X	Public (PU)		Sensitive (SEN)		Classified		
	PU: Public, fully open SEN: Sensitive, limited under the conditions of the Grant Agreement Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444 Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444 Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444							

Legal disclaimer

This project is funded by the European Union under Grant Agreement No. 101083857 and co-funded by the UK Research and Innovation Grant Award No. 10067111. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

DOCUMENT HISTORY OF CHANGES

Version	Date	Author and Short Org. Name	Description
V0.1	2023/09/11	José Barateiro (LNEC)	Initial draft
V0.2	2023/11/04	José Barateiro (LNEC)	DMP with dataset catalogue
V0.3	2023/11/23	José Barateiro, António Antunes (LNEC)	Complete version before internal review
V1.0	2023/11/28	José Barateiro, António Antunes (LNEC)	Complete version including internal reviews

DOCUMENT REVIEW

Reviewer	Date	Reviewer Name (Short Organisation Name)
1	27/11/2023	Dr Vasileios Margaritis (KMOP)
2	27/11/2023	João Palha Fernandes (LNEC)
3	27/11/2023	Margarida Rebelo (LNEC)

ABBREVIATIONS

Abbreviation	Definition
AES	Advanced Encryption Standard
BPMN	Business Process Modelling Notation
DMP	Data Management Plan
DOI	Digital Object Identifiers
EAG	Ethics Advisory Group
ES	Experimental sites
FAIR	Findable, Accessible, Interoperable and Reusable
GDPR	General Data Protection Directive
IPR	Intellectual Property Rights
NBT	Nature-based therapies
TESSA	Toolkit for Ecosystem Service Site-Based Assessment

Table of contents

Index of figures

Index of tables

1. Introduction

The aim of this document is to establish the initial version of the Data Management Plan (DMP) for the NATURELAB project, defining the data management strategy throughout the data lifecycle, as well as identifying and classifying the relevant data used and/or produced in this project.

To manage and control the relevant datasets in this project, a “NATURELAB dataset catalogue” spreadsheet was created which aims to effectively record and characterise the relevant details in the context of the DMP, namely, data on characterization, security, access and applied to each of the Findable, Accessible, Interoperable and Reusable (FAIR) principles.

Findable: Each dataset must be easily found by potential users. To achieve this goal, a unique and persistent identifier will be assigned to each relevant dataset, and comprehensive descriptive and preservation metadata will be generated in accordance with established metadata standards. This metadata will be findable by public services like DataCite.

Accessible: To ensure that datasets can be easily accessed, relevant metadata can be harvested using open protocols such as OAI-PMH and a Web-API (REST services), allowing for secure authentication and authorisation when needed.

Interoperable: Metadata and file formats will adopt standards and/or formal representations with interoperability capabilities. In order to achieve this goal, migration before ingest (to FAIR repositories) will be followed when original formats are closed formats or highly specific, limiting future interoperability.

Reusable: Datasets will be published along with metadata to facilitate their future reuse. To address concerns regarding sensitive data, anonymisation techniques will be applied whenever possible, enabling the publication of datasets without authorisation requirements.

Due to the variety and different nature of data (e.g., data from interviews or data captured by sensors), it is fundamental to provide a simple method where data owners can register all details about datasets in a coherent and normalised way. Therefore, since the “NATURELAB dataset catalogue” is part of the DMP, the DMP is a living document where details about new datasets or updates on existing ones (e.g., new versions) are continuously updated on this catalogue. Also, new versions of the DMP will be reported as formal project deliverables on months M36 and M54, and updates will be tracked on Periodic Project Reports.

This document presents the workflows that should be followed during the data creation and storage processes, which depend on several aspects of specific datasets. For instance, if the dataset handles personal data, informed consents should be used.

Additionally, this deliverable presents a set of strategies related to data security and ethical issues that should be considered in the data lifecycle. For instance, the level of confidentiality can be affected by anonymisation techniques, making it possible to disclose data that could not be made public otherwise.

Finally, each dataset will detail how legal and ethical requirements are managed, ensuring compliance with GDPR and other national regulations that may apply where datasets were produced. Whenever possible, anonymisation techniques are adopted to convert sensitive datasets into anonymised datasets that might possibly be disclosed according to FAIR principles.

The remainder of this document is organized according to the structure defined in the DMP template for the Horizon Europe program, namely: Section 2 - Data summary provides and overview of datasets produced or reused in the scope of this project; Section 3 - FAIR Data, outlines the compliance to FAIR principles, ensuring that datasets are as open as possible, technically normalized and include relevant metadata to be Findable, Accessible, Interoperable and Reusable; Section 4 - Allocation of resources, details potential relevant costs required to make data FAIR; Section 5 - Data security, describes key security considerations, focusing primarily on confidentiality restrictions; and Section 6 refers to relevant ethical aspects in the course of data management.

2. Data summary

NATURELAB aims to increase the recognition, promotion, and use of green and blue spaces (areas with vegetation and water bodies) as care providers. It explores the advantages of so-called "nature-based therapies" (NBT) for people with diverse health needs across various settings. NATURELAB focuses on nature exposure and experiences offered by i) forests and protected areas, ii) urban parks, and iii) horticultural and gardening environments.

As part of the NATURELAB project, nature-based therapeutic programs are being developed, implemented, and evaluated. NATURELAB operates at a total of 15 experimental sites (ES) spread across five countries, including Peru, Portugal, Greece, Germany, and the Netherlands. These countries and locations differ significantly in terms of climate, geography, culture, population density, and healthcare and social care systems. The project will quantify nature exposure and provide nature-based therapies to approximately 4000 participants of all ages, representing diverse socioeconomic backgrounds and levels of health and well-being care needs, including prevention and support for physical (e.g., hypertension) and mental (e.g., depression) conditions. The project will compare long vs. short exposure, active vs. passive activities, type and dose of interactions necessary to achieve health and well-being benefits, and the quantitative and qualitative aspects of natural features (e.g., greenness and presence of water) that are relevant to health and well-being, among other variables.

2.1 Research data

Since the process of identifying and collecting data is still in progress, the initial DMP can only offer a partial representation of the required datasets for the NATURELAB project. The current data summary offers an initial view of the diverse types of datasets essential for the project. To achieve the project's goals, qualitative and quantitative data will be created through several data collection methods and used in each Work Package. NATURELAB will collect data concerning the natural and infrastructural dimensions of each ES, results from the Toolkit for Ecosystem Service Site-Based Assessment (TESSA) method, including noise, daylight/solar radiation and air quality, and statistics and surveys for some of the ES, among others. On the other hand, data will also comprise a variety of data sets, regarding the participants' profile, the psychological health self-assessment or interviewer-administrated psychological health assessment (for participants with mental health disorders with low level of cooperation), and physical measures (e.g., heart rate, blood pressure, blood and urine measurements).

Semi-structured interviews will be conducted to explore the needs and expectations of key stakeholders within and beyond healthcare systems in each country. Audio and other data collected

during the interviews will be collected and stored following the Key Informant Interviews: A Guide for Interviewers [1].

During the project, data will be collected from the field, by completing health and well-being self-assessment questionnaires (or interviewer-administered assessment when needed) and conducting clinical observations. These questionnaires will provide quantitative assessment instruments, namely psychological self-assessment measures (e.g., State-Trait Anxiety Inventory, World Health Organisation Well-Being Index, Comprehensive Geriatric Assessment Toolkit, and Rosenberg Self-esteem Questionnaire), nature assessment (e.g., Dose of Nature) and general physiological measures (e.g., blood pressure, heart rate). Furthermore, blood and urine samples (e.g., indicators of inflammatory responses) will be collected before and after the NBT. Other non-invasive examinations (e.g., bioimpedance measurement, heart rate variability, indirect calorimetry) and a special set of additional questionnaires (e.g., Barratt Impulsiveness Scale, Connor-Davidson Resilience Scale, Food Craving Questionnaire, Sensitivity to Punishment and Reward Questionnaire) will also be used.

Besides clinical observations, each experimental site will be assessed by conducting an on-site characterization and inventory of all nature features (e.g., flora and fauna; water bodies) and infrastructures (e.g., walking trails, seating, and resting facilities). For the six experimental sites located in Portugal, additional characterisation will be done regarding i) exterior daylight and solar radiation environment; ii) sonic context and iii) soundscape characterisation. Site-specific assessment methodologies will be established for this purpose, and specific equipment (supported by data loggers) will be used to measure noise; evaluate human sound perception, and measure daylight. The sites' physical/psychoacoustic and human soundscape perception indicators, as well as the classification of the daylight and solar radiation, will contribute to evaluating the role and significance of these data. Air quality will be monitored at several locations with different levels of exposure to traffic, in Portugal, Peru and the Netherlands, to obtain key air quality indicators (e.g., PM2.5, PM10 and NO2).

2.2 Data Formats

In order to improve future interoperability and seeking for digital preservation, NATURELAB DMP adopts a file format migration before ingest into FAIR repositories. This strategy also intends to reduce the complexity of handling heterogeneous formats during the project execution, ensuring that normalised data formats are adopted as soon as possible throughout the data lifecycle.

Thus, before storing datasets in FAIR repositories, format migration should be adopted if a lossless migration is possible and the original format is not recommended for long-term preservation,

according to recognised recommendations, such as those proposed by *Digital Preservation Coallition*[1].

2.3 Access Levels

The level of confidentiality and access of each dataset is of great importance, as it determines the set of procedures that must be followed in the management of each data set, from the data collection and manipulation to the way in which this data must be stored. In NATURELAB, confidentiality levels are classified in 4 distinct levels.

- Public - data that can be openly accessible to the public.
- Internal - the access to this type of data is limited to the members of the NATURELAB project.
- Confidential - the access to this data is limited to a set of specific people/organisations. When datasets are classified as confidential, the dataset owner must identify the list of people and/or organisations that can access it.
- Secret - confidential data that has sensitive information and must be fully protected against unauthorised access. Clearance can only be provided to specific people and all accesses must be recorded.

2.4 Data owners and access rights

The NATURELAB's grant agreement [2] establishes the Intellectual Property Rights (IPR), access rights and rights of use in its Article 16, further detailed in Annex 5.

Without foregoing a detailed review of the grant agreement, the following aspects are relevant for determining the framework regarding data ownership.

Firstly, it is important to distinguish between: (*i*) background materials (e.g., data, know-how, methods, information) existing before the project, but the project execution requires its be reuse, and (*ii*) data (in all forms considered in this DAM) produced during the project. Naturally, only background materials necessary for the correct execution of the NATURELAB project should be considered as background.

In the case of reused background, it is relevant to consider that the rights to this background may belong to consortium members or to a third-party. The consortium members must identify in a written agreement the background needed to implement the project. On a case-by-case basis, rules for

[1] https://www.dpconline.org/

accessing this background should be agreed upon among the consortium members. If the rights belong to a third-party, it is necessary to ensure compliance with all obligations with that entity for the use of that background.

Concerning data produced during the project, the ownership of this results is not transferred to the granting authority. As stated in the grant agreement [2], results are owned by the consortium members that generate them. However, two or more beneficiaries' own results jointly if: they have jointly generated them and it is not possible to establish the respective contribution of each beneficiary, or separate them for the purpose of applying for, obtaining or maintaining their protection. Under the grant agreement, the joint owners must agree — in writing — on the allocation and terms of exercise of their joint ownership ("joint ownership agreement"), to ensure compliance with their obligations.

Shared ownership is particularly relevant in processed/analysed data where new data/information is potentially produced from multiple sources that can be owned my different consortium members. For these cases, it is important to note that processed/analysed data although considered derived data, it must be managed independently as a new dataset. Therefore, this new dataset falls under the scenario of shared ownership, requiring a "joint ownership agreement" among the involved partners, agreeing on the allocation and terms of exercise of their joint ownership.

2.5 Data management methodology

Figure 1 explains the data management methodology that must be followed in the NATURELAB project, from data creation to storage/archiving. The figure uses Business Process Modelling Notation (BPMN)[2] to represent the general procedures involved in the NATURELAB data management lifecycle.

Firstly, it is important to note that datasets can be produced using different methods. In the context of the NATURELAB project, the main data will be produced through observation/measurement methods, processed/analysed data, semi-structured qualitative interviews and surveys. Data can also be reused (data from previously existing data sources), or produced through other methods, such as collaborative creation.

The data owner, in cooperation with the data manager is responsible to comply with the data management methodology defined in this DMP. Thus, data ownership must be established for each dataset, according to the guidance referred in section 2.4.

[2] https://www.omg.org/spec/BPMN/2.0/

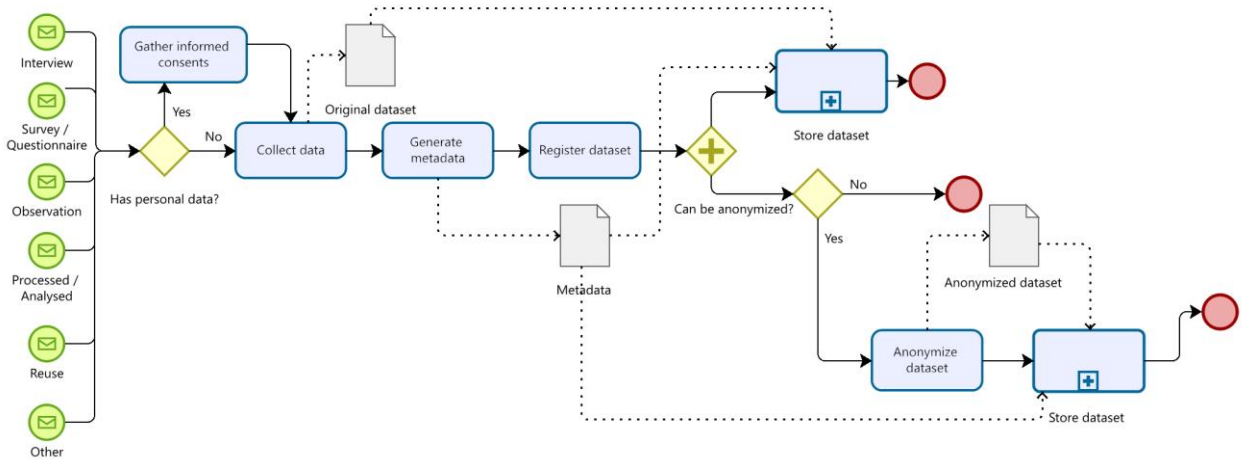

Figure 1 - NATURELAB Data management methodology

Before data is actually collected, it is critical to identify the existence of personal data. If personal data exists, compliance with the legal requirements established by the EU General Data Protection Directive (GDPR) must be ensured and informed consents must be gathered.

Note that the definition of procedures for obtaining informed consent and other legal provisions depends on each data gathering procedure and must be specialised for each case. As an example, in the context of Task 4.1, the procedures for conducting interviews were already specified, which are detailed in the document Key Informant Interviews: A Guide for Interviewers [1], which details all data collection and processing procedures, and defines the informed consents required for this specific type of datasets.

After the data is effectively produced, it is necessary to characterise it to ensure compliance with FAIR requirements. Therefore, it is necessary to generate the relevant metadata and register the dataset in the “NATURELAB dataset catalogue”.

The “NATURELAB dataset catalogue” is organised in 6 sections, namely:

- Data Summary: general overview of the dataset, including the characterisation and description of its content, and its purpose and relation to the project;
- FAIR - Findable: details about metadata provisions and conventions to ensure that the dataset is discoverable with appropriate metadata;
- FAIR - Accessible: details about the access level of the dataset and justifications for restricted access when the dataset cannot be made public;

- FAIR - Interoperable: details about technical interoperability, including formats and vocabularies;
- FAIR - Reusable: details to support future data reuse, including documentation and quality assurance to trust the produced data;
- Ethics and Legal Aspects: details about ethical and legal aspects that must be covered by each dataset.

Note that the components related to FAIR principles are detailed in section 3, and details about ethical and legal aspects are detailed in section 7.

After generating the relevant metadata and registering the dataset in the NATURELAB dataset catalogue, the dataset must be stored.

When datasets are anonymised, using anonymization techniques that ensure an irreversible process, an additional dataset is created, and two datasets must be handled: the original dataset and the anonymised dataset. From this point, the data will be treated in the same way, depending only on the level of confidentiality. For example, it is expected that the original data will have a high level of confidentiality and that anonymised data may be public.

Figure 2 details the storage process, which depends on the level of confidentiality.

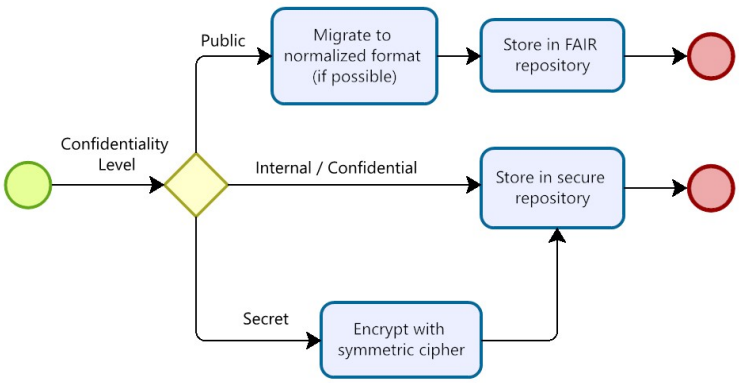

Figure 2 - Store dataset process

In the case of public datasets, if the original format is not adequate for long-term preservation and it is possible to loosely migrate the dataset to a format suitable for long-term preservation (as indicated in section 3.3), the dataset should be migrated to that format, before ingest. Then, the dataset must be deposited in an open repository that complies with FAIR requirements.

This DMP does not restrict the repository to be used. Community, national, institutional, or generic repositories such as Zenodo[3] can be used, as long as they are FAIR compliant. For instance, data produced in the scope of interviews detailed in Key Informant Interviews: A Guide for Interviewers [1], that can be made public, will be stored in the VU Yoda[4] repository.

All datasets that are not public (Internal, Confidential and Secret) must be stored in a controlled and secure repository. Due to their level of sensitivity, secret datasets must always be encrypted using a secure symmetric cipher (the use of AES - Advanced Encryption Standard with 256-bit keys is recommended). Key management must be defined by the dataset owners.

As previously mentioned, data produced in the scope of interviews will be securely stored on the VU data storage platform Research Drive[5]. Similarly, this DMP does not restrict the adoption of a single secure storage. Specific datasets might be stored in other secure repositories as long as they comply with the above-mentioned process to ensure Confidentiality, Integrity and Availability of datasets.

2.6 NATURELAB Dataset Catalogue: Data Summary

The NATURELAB data catalogue: data summary is organised by the following topics:

- **Data ID**
  # (sequence number) that uniquely identifies each dataset
  Example: 1, 2, 3, …

- **Name**
  Name of the data set. This is an open field, and no naming convention is defined[6].
  Example: NATUREALAB Stakeholders Mapping.

- **Date**
  Date of creation and / or reuse of the dataset: This field details the date of each particular dataset. If the dataset had multiple versions, this field always contains the date of creation, i.e., it is the same for all versions. It is not mandatory that the level of detail for this field is on a day level. For instance, October 2023 is an acceptable date.
  Example: October 2023.

- **Version**
  Version number: This is a sequential number. Note that the same dataset can have multiple versions. In that case, DATA ID must be the same. The pair (Data ID, Version) uniquely identifies each dataset.
  Example: 1, 2, 3.

[3] https://zenodo.org/
[4] https://yoda.vu.nl/site/
[5] https://vu.nl/en/research/storagefinder
[6] The DMP does not force any naming convention as specific fields are captured in the catalogue register. Thus, derived names composed, for instance, by work package, experimental site, version, can always be generated from the structured information.

- **Version Date**
  Date of this particular version: This field details the date of this particular version. If the dataset has only one version, this field contains the same value of the "Date" field. If it is, for instance, the second version of the dataset, this field contains the date of the second version, while the "Date" field contains the date of creation of the original (first) version.
  Example: November 2023.

- **Version Description**
  Open field to describe this version of the dataset. The dataset owner must provide details about the differences from the previous versions, explaining Why this is a new version.
  Example: First version of the dataset.

- **Type**
  Type of the dataset. As a first level of classification the type can be Digital or Physical. It is expected that all datasets produced in NATURELAB will be Digital datasets. The second level of classification further details the type of the dataset. This is a normalised field where values can be: Digital (data, software, workflow, protocol, model, other); Physical (material, reagent, other). This list can be extended during the project, if new relevant types are identified.
  Example: Digital: data

- **Type of data collection**
  When applicable, identifies the method used for data collection. For instance, experimental, observational, analysed/processed data, meta-analysis, surveys.
  Example: Observational data

- **Authors**
  List of authors/contributors to the creation of the dataset. The authors can be specific people or organisations.
  Example: All partners of the project consortium

- **Ownership**
  Identify the owners of this dataset. If the ownership is shared, provide details about the joint ownership agreement.
  Example: All partners of the project consortium. Equitable share among partners.

- **Location (repository)**
  Location of the dataset in a specific repository, such as Basecamp, Yoda, Research Drive, Google Drive, specific information system. If the dataset is located in more than one repository, all locations must be identified.
  Example: Basecamp/WP5folder

- **Generated / Reused**
  Identifies if the dataset was generated (G) in NATURELAB or reused (R) from previous works.
  Example: G

- **Purpose and relation to project objectives**
  Open field to detail the purpose of this dataset and how it relates to the project objectives, explaining the reasons to reuse or to generate this data for NATURELAB.
  Example: To allow systematic approach towards establishing strong links with all relevant stakeholders.

- **Experimental sites**
  If applicable, identify the experimental sites related to this dataset.
  Example: All

- **Work Packages**
  Details how Work Packages are related to this specific dataset, providing an overview of how this particular dataset relates to the overall project structure. The Work Package mapping must be classified as Consume (C - for Work Packages that consume/read this dataset) or Produce (P - for Work Packages that produce data for this dataset). Note that due to the multiple activities developed by each Work Package, the same Work Package can both produce and consume the dataset.
  Example: P: WP5; C: WP5, WP5

- **Target users**
  Identifies the potential users/stakeholders interested in this dataset. The identification of target users must cover the period during and after project.
  Example: Project partners; Social Innovation Hub; sister projects

3. FAIR data

The open data management of NATURELAB is based on the quality control of the research-related resources, according to the principles of FAIR, and “as open as possible, as closed as necessary”. Indeed, datasets that do not include sensitive data will be published for open access. Anonymisation and pseudo-anonymisation techniques are applied to make it possible to also publish datasets that include sensitive information. Secure management of datasets will be ensured, and all the principles of the General Data Protection Regulation and specific national legislation will be safeguarded.

3.1 Making data findable, including provisions for metadata

All open data, publications and open-source software produced during the NATURELAB project must be easily identifiable and findable by potential users (inside and outside of the project). Persistent and unique identifiers, such as Digital Object Identifiers (DOI[7]), should be created and assigned to each public dataset and publications.

DOI assignment will be performed automatically by the trusted repository used for specific datasets. Note that this DMP allows the use of multiple trusted repositories, from community based, to organizational, national, or generic ones. As a last resort, a Zenodo project space will be used. Note that decisions about specific repositories will only be made as the data is being created. Currently, in the scope of Task 4.1, Key Informant Interviews: A Guide for Interviewers [1], it is already defined that the Yoda trusted repository will be used to store public datasets resulting from the interviews analysis, while Research Drive[8] will be used to store confidential data produced in the scope of these interviews.

Metadata schemas are used to provide descriptive, technical and preservation metadata, facilitating the search and access to data. The use of known metadata schemas, such as Dublin Core[9] or METS[10], is recommended during the project, making it easier for data to be findable by public services like DataCite. As previously mentioned for DOI generation, the supported metadata schemas can also depend on the trusted repository.

Metadata concerning each dataset will be collected using the NATURELAB dataset catalogue, including a set of keywords used to characterize the dataset and facilitate dataset discovery.

[7] https://www.doi.org/
[8] https://vu.nl/en/research/storagefinder
[9] https://www.dublincore.org/
[10] https://www.loc.gov/standards/mets/

The “NATURELAB data catalogue: FAIR - Findable” section is organised by the following topics:

- **Persistent Identifier?**
  Does the dataset have a persistent identifier? If yes, identify it.
  Example: No

- **Metadata schemas**
  List of descriptive metadata schemas used (e.g., Dublin Core, METS) used to allow dataset discovery. If no metadata schema is used, outline what type of metadata will be created and how.
  Example: N/A

- **Keywords**
  List of keywords that characterise the dataset.
  Example: Stakeholders mapping

3.2 Making data accessible

Data should be stored according to its access level, as shown in Figure 2. Public datasets should be stored in open repositories, following FAIR principles, such as Zenodo[11] or institutional repositories (e.g., Yoda[12]). These repositories usually require a set of descriptive metadata for each dataset, ensuring its findability and accessibility. Data should be stored in a normalised format (see Section 3.3). Open software and data-analysis scripts developed during the project should be made available through repositories such as GitHub.

Internal, confidential, and secret data should be stored in a secure repository, such as Research Drive[13]. Due to its sensitivity level, secret data must be encrypted using a symmetric cipher before storage. Regardless of the level of access, a retention period must be defined for each data set, indicating how long the data will be available.

The “NATURELAB data catalogue: FAIR - Accessible” section is organized by the following topic:

- **Open repository**
  If it exists, identify the OPEN repository (e.g., Zenodo, Yoda, institutional repository)
  Example: N/A

- **Access level**
  Public (P) - data that is openly accessible to the public; Internal (I) - access to this type of data is limited to the members of the NATURELAB project, Confidential (C) - access to this data is limited to a group of people/organisations. When datasets are classified as confidential, the dataset owner must identify the list of people and/or organisations that can access it, Secret (S) - confidential data that contains sensitive information and must be fully protected against unauthorised access. Authorization can only be granted to specific people and all accesses must be recorded. If the dataset is not classified as Public, explain why,

[11] https://zenodo.org/
[12] https://yoda.vu.nl/site/
[13] https://vu.nl/en/research/storagefinder

clearly separating legal and contractual reasons from intentional restrictions. For Confidential and Secret information identify the list of people with authorised access.
Example: I. Consortium only because it may contain personal information (e.g., name; email)

- **Descriptive metadata available?**
  Will metadata be available for any user (Y/N)? If the metadata cannot be disclosed, explain why.
  Example: Yes

- **Retention period**
  For how long will the data be available? Note that this retention period can depend on the choice of the repository. For instance, when using Zenodo, the retention period is set to at least 20 years[14].
  Example: 20 years

3.3 Making data interoperable

Metadata schemas should be represented and encoded using standard data representation such as JSON or XML, which are used by most FAIR repositories and metadata search engines. Table 1 provides an overview of recommended file formats to be used when datasets are stored in FAIR repositories. To the date of this deliverable, since the project will still produce several datasets, this table should be seen as the initial prediction or file formats, being only an initial recommendation. Information about the software required for data renderisation and approximate size of the dataset, including number of records, is also collected in the NATURELAB dataset catalogue.

Table 1 - Recommended file formats

Type of data	File format
Documents	PDF/A (.pdf), Microsoft Word (.docx)
Structured text	XML (.xml)
Tabular data	Comma Separated Values (.csv), Microsoft Excel (.xlsx)
Statistical Data	Comma Separated Values (.csv), SPSS (.por, .sav)
Geospatial	Geographic Markup Language (.gml), GeoTIFF (.tiff), GeoJSON (.geojson)
Databases	SQL scripts using DDL (.sql), SQL scripts for schema creation + Comma Separated Values (.csv) for table records
Images	TIFF (.tiff), PNG (.png), JPEG2000
Audio	WAVE (.wav), MPEG Audio Layer III (.mp3)
Video	Audio Video Interleaved (.avi), MPEG-4 (mp4)

[14] https://about.zenodo.org/policies/

The “NATURELAB data catalogue: FAIR - Interoperability” section is organised by the following topics:

- **Metadata vocabularies (domain specific)**
  If applicable, identify the data and/or metadata vocabularies used in this particular dataset.
  Example: This represents the vocabulary used in the Stakeholders mapping template (e.g.: type of stakeholders)

- **File formats**
  Detail the technical file format used.
  Example: Excel (xlsx)

- **Software required to render**
  Software required to render the dataset.
  Example: MS Excel

- **Size**
  Approximate size of the dataset. When applicable, the approximate size must be identified for both the binary size and the number of instances.
  Example: ~250 KB; approx. 400 records

3.4 Increase data reuse

According to the Grant Agreement, open results developed during the NATURELAB project should be made public under a Creative Commons license (CC-BY-SA or CC-BY), preferably the latest version[15], except in special circumstances that call for a more restrictive type of CC license. Datasets should be documented to facilitate data reuse, following FAIR principles, ensuring, when possible, that metadata is machine actionable.

Embargo periods may apply due to several reasons, such as to publish and disseminate results or seek patents. When embargo periods are applied, dataset owners mut specify why and for how long the embargo applies, considering thar research data should be made available as soon as possible.

The “NATURELAB data catalogue: FAIR - Reuse” section is organised by the following topics:

- **Data documentation**
  Documentation (explanatory text & images to illustrate data generation, readme files, details about compilation processes in case of software) to validate data analysis and facilitate data reuse.
  Example: N/A

- **Provenance data?**
  Data provenance is documented? Any standard is used?
  Example: N/A

[15] http://opendefinition.org/licenses/cc-by/

- **Data quality processes**
  Was the data verified with data quality processes? If yes, which procedures were taken?
  Example: Yes (manual during data compilation)

- **Embargo period**
  Is an embargo period required for this dataset? If yes, define how long is the period and justify the need for the embargo, considering that research data should be made available as soon as possible.
  Example: No

4. Other research outputs

At the time of this initial data management plan, there are no expected research outputs other than software (data-analysis scripts), datasets and publications. However, if any other digital research output is created or obtained during the NATURELAB project, OPEN and FAIR principles will be applied according to the nature of that specific output.

5. Allocation of resources

The NATURELAB DMP guides the effective management of datasets throughout the project's lifecycle, based on resources required for research data quality, FAIR compliance, and maximum openness. The plan ensures adequate resourcing through a dedicated data management task, highlighting the use of repositories like Yoda and Zenodo which do not have additional costs to researchers. Currently, there are no immediate costs anticipated to ensure the FAIRness of data, and any potential costs related to open access publications are eligible for reimbursement according to the grant conditions.

The costs related to human resources allocated to the overall data management in NATURELAB are already planned in the project budget.

The DMP underscores the project's commitment to transparency and accessibility, noting that potential costs will be addressed in future versions of the plan or other relevant project documentation, if necessary.

6. Data security

At the time of this initial data management plan, there are no expected research outputs other than software (data-analysis scripts), datasets and publications. However, if any other digital research output is created or obtained during the NATURELAB project, OPEN and FAIR principles will be applied in accordance with the nature of these specific outputs.

The security provisions in NATURELAB strictly depend on the level of confidentiality of each dataset, namely:

- Public - data that can be openly accessible to the public;
- Internal - access to this type of data is limited to the members of the NATURELAB project;
- Confidential - access to this data is limited to a specific set of people/organisations. When datasets are classified as confidential, the dataset owner must identify the list of people and/or organisations that can access them;
- Secret - confidential data that has sensitive information and must be fully-protected against unauthorised access. Authorization can only be granted to specific people and all accesses must be recorded.

The following table details the security procedures that must be followed for each confidentiality level.

Table 2 - Security procedures

Actions	Public	Internal	Confidential	Secret
Printing	No restrictions	Printed copies must be recorded and destroyed after usage	Printed copies must be recorded and destroyed after usage	Printing is not allowed for Secret data
E-mail	No restrictions	Reference to the document location can be sent by e-mail, using project mailing lists or directly to project participants. The document itself cannot be attached to the e-mail	Reference to the document location can be sent by direct e-mail to authorised users. The document itself cannot be attached to the e-mail	Reference to the document location can be sent by a direct e-mail to authorised users. The document itself cannot be attached to the e-mail. Encryption keys must be shared using a

Actions	Public	Internal	Confidential	Secret
				distinct channel (other than e-mail)
Distribution	No restrictions	Allowed within the project consortium	Must be kept to a minimum set of authorised users.	Must be kept to a minimum and numbered distribution lists must be kept
Local copies	No restrictions	Local copies can be stored in personal computers with authentication mechanisms	Local copies can be stored in personal computers with authentication mechanisms	Local copies must be encrypted using secure symmetric ciphers, such as AES. Encryption keys must be managed outside the scope of the local storage
Storing Material	Public open repositories	Project shared folders, e.g., Basecamp	Secure repositories	Encrypted in secure repositories

Considering the common dimensions of information security, Confidentiality, Integrity and Availability, the above-mentioned procedures are used to ensure the confidentiality of information.

Integrity and availability are delegated to trusted repositories as soon as datasets are ingested into them. These repositories use standard techniques to ensure bitstream integrity (redundancy and checksums) and also provide backup and redundancy mechanisms to ensure availability.

During dataset creation, dataset owners are responsible to ensure data integrity and availability. Although not mandatory, users should perform regular backups of their data.

For datasets centrally managed, e.g., by applications supporting experimental sites monitoring or collecting data from surveys, a backup plan must be defined, complying, at least, with the following frequency: total backups once per month; differential backups once a week and incremental backups once a day.

Finally, source code (e.g., scripts for data analysis) must be committed to a versioned source-code repository. This DMP recommends the use of of git[16] based repositories, such as github[17].

[16] https://git-scm.com/
[17] https://github.com/

7. Ethics

The NATURELAB project has established an Ethics Advisory Group (EAG), tasked with offering guidance on ethical, legal, and societal matters that may arise throughout the project. The EAG ensures that data collection follows international guidelines. Each experiment will only begin after receiving local/national ethical review board or committee approval, as mandated by applicable laws or regulations. Deliverable D6.2 NATURELAB Ethics Guidelines [3], which includes templates for project information sheet, certificate of consent for adult participants, certificate of consent for parents/guardians, certificate of assent for children, confidentiality agreement and media waiver, establishes the NATURELAB approach to ethics and provide specific guidelines for data collection and authorisation requests in distinct countries, namely, Germany, Greece, Peru, Portugal and The Netherlands.

Specific procedures are aligned with this DMP, according to the methods detailed in this deliverable. Specific information related with legal and ethical aspects must be registered in the specific section of the "NATURELAB dataset catalogue - Ethics and Legal Aspects", which includes the following topics:

- **Is the dataset dealing with personal data?**
  Yes/No
- **If personal data is involved, do we have an informed consent for data sharing and preservation?**
  Yes/No
- **Is the dataset anonymised?**
  Yes/No
- **Project information sheet**
  Relevant and understandable information about the project was shared, including on how the data will be used and for what purpose.
- **Consent form**
  Appropriate procedures for pertaining consent have been followed (written form/ oral consent, pertinent to subject's characteristics…), as well as concerns sharing, retention, usage, and deletion of data.
- **Institutional ethics review**
  If an ethical review procedure is required by the institution where the consortium partner works at, this procedure has been adequately followed.

References

[1] Task 4.1 - Key Informant Interviews: A Guide for Interviewers, VU, October 2023.

[2] Project 101083857 Grant Agreement – NATURELAB - HORIZON-CL6-2022-COMMUNITIES-02-two-stage, May 2023.

[3]. D6.2 – NATURELAB Ethics Guidelines, November 2023

Appendix 1 – NATURELAB dataset catalogue

Section	Attribute	Description	Example
Data Summary	**Data ID**	*# (sequence number)*	1
	Name	*Data name*	NATURELAB Stakeholders Mapping
	Date	*Date of creation / reuse of the complete data set*	October 2023
	Version	*Version number.*	1
	Version Date	*Date of versioning*	October 2023
	Version Description	*Details about this specific version*	First version
	Type	*Data type (digital / physical)* *Digital: data, software, workflow, protocol, model, etc.* *Physical: new material, reagent, etc.*	Digital: data
	Type of data collection	*If applicable, name the type of data collection used (e.g., experimental, observational, analysed/processed data, meta-analysis)*	
	Authors	*List of authors / contributors (people and/or organisations)*	All partners
	Location (project repository)	*Location in project repository, e.g., Basecamp, Yoda, Research Drive, data storage platform, information systems, etc*	Basecamp/ WP5 folder
	Generated / Reused	*Generated within the project (G) or reused (R)*	G
	Purpose and relation to project objectives	*Describe the data and related it with project objectives*	To allow systematic approach towards establishing strong links with all relevant stakeholders
	Experimental Sites	*Identify experimental sites related to this data*	All
	Work Packages	*Related WP to consume (C) and/or produce (P) data*	P: WP5 C: WP4; WP5
	Target users	*Potential users interested in this data (during and after project)*	Project partners; Social Innovation Hub; sister projects
Fair - Findable	**Persistent Identifier?**	*Does the data have a persistent identifier? If yes, identify it.*	No
	Metadata schemas	*List of metadata schemas used (e.g., dublin core, mets)*	N/A
	Keywords	*List of keywords that characterize the dataset*	Stakeholders mapping
Fair - Accessible	**Open repository**	*If exist, identify the OPEN repository (e.g., Zenodo, institutional repository)*	N/A
	Access level	*Public (P), Internal (I), Confidential (C) or Secret (S). If not public, explain why*	R. Consortium only because may contain personnel info (e.g. name; email)
	Descriptive metadata available?	*Will metadata be available for any user (Y/N)?*	Yes
	Retention period	*For how long will the data be available?*	To be defined (TBD)

Fair - Interoperable	**Metadata vocabularies (domain specific)**	*Vocabularies to describe data*	This represents the vocabulary used in the Stakeholders mapping template (e.g.: type of stakeholders)
	File formats	*File formats used*	Excel (xlsx)
	Software required to render	*Software required to render the data*	Excel
	Size	*Aprox size of the dataset (number of instances #; and binary size)*	~250 KB; approx. 400 records
Fair - Reusable	**Data documentation**	*Documentation (explanatory text & images to illustrate data generation) to validate data analysis and facilitate data reuse*	N/A
	Provenance data?	*Data provenance is documented? Any standard is used?*	N/A
	Data quality processes	*Was the data verified with data quality processes? If yes, which procedures were taken?*	Yes (manual during data compilation)
Ethics and Legal Aspects	**Is the dataset dealing with personal data?**	*Yes/No*	
	If personal data is involved, do we have an informed consent for data sharing and preservation?	*Yes/No*	
	Is the dataset anonymised?	*Yes/No*	
	Project information sheet	*Relevant and understandable information about the project was shared, including on how data will be used and what for.*	
	Consent form	*The appropriate procedures for pertaining consent have been followed (written form/ oral consent, pertinent to subject's characteristics…), and consent also concerns sharing, retention, usage and deletion of data.*	
	Institutional ethics review	*In case going through an ethical review procedure is required by the institution that consortium partner at hand works at, this procedure has been adequately followed.*	

NATURELAB

www.naturelab-project.eu