



Deliverable 1.1

Data Management Plan – June 20 2024

BEPREP has received funding from the European Union (Grant agreement No 101060568).



**Funded by
the European Union**

Program	HORIZON EUROPE
Call	HORIZON-CL6-2021-BIODIV-01
Project Name	Identification of best practices for biodiversity recovery and public health interventions to prevent future epidemics and pandemics
Project Acronym	BEPREP
Project Number	101060568
Project Start / Duration	1 Sep 2022/54 months
Project Coordinator	University of Helsinki
Beneficiary in Charge of the Deliverable	Swedish University of Agricultural Sciences (SLU)

Dissemination Level		
PU	Fully open (automatically posted online on the Project Results platforms)	X
SEN	Limited under the conditions of the Grant Agreement	
RESTREINT-UE/EU-RESTRICTED	under Decision 2015/444	
CONFIDENTIEL-UE/EU-	under Decision 2015/444	
SECRET-UE/EU-SECRET	under Decision 2015/444	

Deliverable Information

Document Administrative Information	
Deliverable Number	1.1
Deliverable Title	Data Management Plan
Beneficiary in Charge	Swedish University of Agricultural Sciences
Contractual Date	19/07/2024
Report Submission Date	05/07/2024
Nature of the Deliverable	R
Lead Author(s)	Hussein Khalil
Contributor(s)	Frauke Ecke, Nivison Nery jr

Change Log

Date	Version	Author/Editor	Summary of Changes
25/06/2024	2	Hussein Khalil	Added missing info re: authorship and contributors, details on current status of DMP, edits to FAIR data and data utility

BEPREP Executive Summary

The BEPREP project will adhere to the "FAIR" principles—Findable, Accessible, Interoperable, and Re-usable—in managing its data. This approach ensures that data will be made findable and accessible within the consortium and to the broader research community, stakeholders, and policymakers. The data management practices will comply with national and European ethical and legal frameworks, including the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679).

This data management plan (DMP) outlines the data management lifecycle for all data collected, processed, and/or generated by the BEPREP project. It provides comprehensive details on the handling of research data during and after the project's duration. Key aspects covered include the nature of the data, methodologies and standards applied, data sharing and open access provisions, and data curation and preservation strategies.

This initial version of the DMP will be continuously updated as the project progresses, incorporating measures taken by the partners to ensure FAIR data principles. An updated version of this DMP will be submitted by 30 June 2025. A final version of the DMP will be delivered at the end of the project (M48, December 2027).

Table of contents

BEPREP has received funding from the European Union (grant agreement No 101060568).

1. BEPREP Executive Summary	3
2. Table of contents	4
3. Abbreviations.....	5
4. Data information.....	5
Data Summary	5
Re-use of Data.....	7
Types and Formats	7
Data Purpose.....	8
5. Data utility, security, and FAIR data.....	10
Data Utility.....	10
Data Security.....	10
FAIR Data	10
Making Data Findable, Including Provisions for Metadata.....	10
Making Data Accessible.....	11
Making Data Interoperable.....	11
Increasing Data Re-use	11
6. Ethics	12
Collection of personal data	12
Collection of Animal Samples	12
Collection of environmental data	12
Exchange of biological samples between EU and non-EU countries	13
Ethical management and oversight.....	13
Compliance with horizon Europe ethical standards	13
Deliverables in WP7 on ethics	13
Monitoring and reporting.....	13

Abbreviations

ABS: Access and Benefit Sharing
 BACI: Before-After-Control-Impact
 BEPREP: Biodiversity and Pathogen Research Project (assumed meaning based on context)
 CSA: Case Study Area
 ECDC: European Centre for Disease Prevention and Control
 EFSA: European Food Safety Authority
 ELISA: Enzyme-Linked Immunosorbent Assay
 ENA: European Nucleotide Archive
 EPI: Epidemiological
 EPI: Epidemiological
 ETC/BD: European Topic Centre on Biological Diversity
 FAIR: Findable, Accessible, Interoperable, and Reusable
 GBIF: Global Biodiversity Information Facility
 GDPR: General Data Protection Regulation
 GIS: Geographic Information System
 IPBES: Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services
 ITS: Internal Transcribed Spacer
 NA: Not Applicable
 NCBI: National Center for Biotechnology Information
 PCR: Polymerase Chain Reaction
 PID: Persistent Identifier
 RED: Research Electronic Data Capture
 SRA: Sequence Read Archive
 SSL/TLS: Secure Sockets Layer / Transport Layer Security
 UANT: University of Antwerp (assumed meaning based on context)
 WHO: World Health Organization
 WP: Work Package

Data information

Data Summary

BEPREP aims to understand the impact of restoration efforts and nature-based interventions on zoonotic risk by studying their effects on biodiversity and species interactions and pathogen transmission. The project includes 12 case study areas (CSAs), and the results and deliverables will be primarily based on empirically collected data that is analysed and modelled during the project. The main types of data are presented in table 1.

Table 1. Main types of data used in BEPREP and their storage formats

Data Type	Description	Storage Formats
Field-collected environmental, ecological, and epidemiological data	Quantitative and qualitative data	Spreadsheets, REDCap database, geo-located
GIS and remote-sensed data	Drone flights, satellite images	Shapefiles, images
Pathogen data	Quantitative data	FASTA, NeXML

Data Type	Description	Storage Formats
Human data	Aggregated incidence data and individual-level data from brazil and tanzania	Spreadsheets, REDCap database

The environmental and animal samples collected in the 12 CSAs will be analysed for a range of zoonotic and potentially zoonotic pathogens (disease X). Biological samples from human participants in Sweden, Brazil, and Tanzania will be tested for important zoonotic pathogens, including orthohantaviruses, arboviruses, and bacterial infections such as leptospirosis and enteric bacteria.

Table 2. A summary of the environmental, animal-related, and human data that will be collected in each case study area (CSA).

CSA	Environmental data	Animal data	Human data
Sweden (CSA 1)	Soil, water, sediment, habitat and remote-sensed surveys	Small mammal, beavers, Reindeer, mosquitoes/ticks	*pending: Hunters/control group (cross-sectional sero-survey), Reindeer herders (questionnaires)
UK Forth (CSA 2)	Soil, water, sediment, habitat surveys	Small mammals, mosquitoes/ticks	NA
Germany (CSA 3) – *Pending initiation of sampling	Soil, water	Small mammals, mosquitoes/ticks	NA
France (CSA 4)	Soil, water, sediment, habitat surveys	Small mammals, bats, mosquitoes/ticks	NA
Italy (CSA 5)	Soil, water, sediment, habitat surveys	Small mammals, bats, mosquitoes/ticks	Perception and habitat use data in study areas
Brazil (CSA 6)	Soil, water, habitat surveys	Small mammals, bats, invertebrates, mosquitoes/fleas	Individual data (residents), longitudinal open cohort design (sero-surveys every 6 months)
Colombia (CSA 7)	Soil, water, habitat surveys	Small mammals, bats, dogs, horses, mosquitoes/ticks	*pending: In-depth interviews, questionnaires of farmers and keepers of livestock
Tanzania (CSA 8)	Soil, water	Small mammals, bats, mosquitoes/ticks	*pending: Individual data (residents), longitudinal open cohort design (sero-surveys every 6 months)
Madagascar (CSA 9)	Soil, water, sediment, habitat surveys	Small mammals, lemurs, mosquitoes/ticks	NA

CSA	Environmental data	Animal data	Human data
Congo (CSA 10)	*pending	*pending	NA
Malaysia (CSA 11)	Soil, water, habitat surveys	Small mammals, bats, monkeys, mosquitoes/ticks	NA
UK – London (CSA 12)	Soil, water, habitat surveys	Small mammals, mosquitoes/ticks	NA

Re-use of Data

Throughout the lifetime of the project, BEPREP will combine data collected through project activities across the WPs (especially WPs 3 and 5) and publicly available ecological, environmental/meteorological, epidemiological, and other monitoring data made available by national and international organizations and institutions. The re-use of data will be for:

- **Understanding Contexts Pre-Restoration:** Utilise previously collected monitoring data within or near the study area and region to improve our understanding of environmental, animal, and public health contexts before restoration efforts, in cases where the interventions have already been completed or initiated.
- **Informing Early-Stage Modelling:** Use pre-existing data to inform initial modelling exercises, allowing the adaptation of the model at later stages when newly generated data will update and improve performance. This is especially relevant for WPs 2 and 6.
- **Determining Sample Sizes:** Use pre-existing data to help in power calculations to determine necessary sample sizes in different case study areas. This includes calculating required sample sizes of small mammals based on previous abundance/density estimates and determining sample sizes needed to estimate pathogen prevalence in vectors (e.g., ticks and mosquitoes), animals (e.g., rodents), and the environment (e.g., water and soil), based on previously estimated prevalence. See table 3 for current sample sizes.
- **Human Data in Four Case Study Areas:** In the case study areas where BEPREP works (or will work) with human data (Sweden, Brazil, Tanzania, Colombia, and Italy), publicly available reporting data on the incidence of various zoonotic pathogens will be utilized. This includes arboviruses in Brazil, leptospirosis and malaria in Tanzania, tularemia and nephropathia epidemica in Sweden, and tick and mosquito-borne infections in Italy. These data will help inform models, identify high-risk areas and periods, and understand the burden of disease.
- **Phylogenetic Data:** Use existing phylogenetic data on both targeted and non-targeted pathogens to inform on potential reservoirs and vectors of these pathogens. This data will aid in assessing the risk of future outbreaks that may be caused by previously unidentified pathogens with zoonotic potential. These data will be available through publicly accessible databases and repositories (e.g., ENA, GISAID, VIPR, NCBI).

Types and Formats

Our data and research output from this project will include:

- **Quantitative BACI Data:** Before-After-Control-Impact data to evaluate the effects of interventions.

- Observational Data: Collected across spatial gradients.
- Multimedia Outputs: Images, audio, and video output from community engagement activities.
- Georeferenced Remote Sensing Data: Images from drone flights and other remote sensing activities.
- Amplicon Sequence Data: 16S and ITS sequence data to study microbial communities.
- Gene Expression and Pathogen Abundance Data: Information on gene expression levels and pathogen quantities.
- Publicly Available Geospatial Data: Combined with our observational and experimental data.
- Experimental Data: Collected by partner projects.

The scale of the data ranges from local (field sampling, observational, audio-visual) to regional (satellite imagery) to national and global (surveillance data, species distributions, environmental variables).

Data Purpose

BEPREP data and research outputs include quantitative BACI (before-after-case-control) data, observational data across spatial gradients. We also gather georeferenced images from remote sensing activities, such as drone flights. Additionally, we obtain large amounts sequence data and information on gene expression and pathogen abundance. These data are combined with publicly available geospatial data, and observational and experimental data collected by partner projects.

The scale of the data ranges from local (e.g., field sampling, observational, audio-visual) to regional (e.g., satellite imagery) to national and global scales (e.g., surveillance data, species distributions, diversity, environmental and climatic variables). Our research outputs consist of curated publicly available datasets, high-resolution raster files for risk predictions (model outputs), reservoir/vector distributions, and environmental characteristics. We also produce statistical packages for open access, along with associated documentation on GitHub.

Specifically, the data generated and re-used by our consortium falls into four categories: environmental data, animal data, pathogen data, and human data.

Environmental data are generated mainly from soil, water, and sediment samples, as well as through habitat and landscape surveys. These samples are analysed for microbiota, pathogen presence, and environmental DNA (eDNA) to identify various organisms in the environment. The habitat and landscape data are collected through on-site surveys, satellite images, and drone flights, stored in spreadsheets and as geo-referenced images and shape files, providing information on habitat characteristics, topography, and climate variables.

Animal data are collected through live trapping of potential reservoirs and terminal sampling of vectors (such as ticks, mosquitoes, and fleas). This data are broadly classified into camera trap data and live trapping data. Our consortium actively samples rodents, birds, and bats to record demographic characteristics, with blood and faecal/urine samples taken for pathogen testing. This data are georeferenced and combined with habitat and landscape data from study areas. Camera traps and faecal samples are used to quantify biodiversity and the presence and abundance of mammals and birds. Ticks and fleas are collected from live-trapped animals and field sampling (e.g., drag sampling), while mosquitoes are sampled through "ovitraps". These data are coupled with environmental data, such as temperature, humidity, and habitat structure.

Pathogen data are obtained by testing environmental (water, sediment, soil) and animal (blood, feces, saliva, tissue) samples for targeted and non-targeted pathogens. In Brazil and Tanzania, human blood samples are also tested for important zoonotic pathogens. Environmental and animal samples undergo phylogenetic analyses and testing for unknown pathogens with zoonotic potential. The data generated are quantitative, stored in spreadsheets that combine individual (e.g., demographic) and habitat/landscape data. PCR is used to detect specific pathogen DNA or RNA in these samples, while ELISA is used to detect and quantify antibodies, providing information on pathogen exposure.

Human data collected and re-used by our consortium are diverse, ranging from aggregated incidence or case data retrieved from public health agencies (e.g., Sweden, Italy) to newly collected individual-level quantitative and qualitative data (e.g., Tanzania, Brazil). In Brazil and Tanzania, quantitative data are collected through periodic sero-surveys using multiple-choice questionnaires, and qualitative data through in-depth interviews and focus groups, which include notes and audio recordings. In Colombia and Italy, we will implement questionnaires to understand how participants perceive restoration efforts and how they use restored and unrestored areas. The questionnaires also cover variables such as gender, age, medical history, general behaviour. These data are collected using a combination of digital and analog means, stored in spreadsheets, and combined with pathogen data following laboratory analyses, including geographic coordinates of households.

By collecting and analysing data from these various sources, our consortium aims to gain a comprehensive understanding of the factors contributing to the emergence and transmission of zoonotic diseases and how restoration and biodiversity efforts impact disease risk in humans.

The data collected for this project enable us to understand how restoration and public health interventions affect disease risk, particularly through changes in biodiversity and ecosystem functioning. We summarize the main purposes for each class of data collected in BEPREP:

- **Environmental Data:** Water, soil, and sediment samples provide data on pathogen contamination, chemical and physical properties (and water quality), and invertebrate and microbiotic diversity. Meteorological data help understand the relationship between weather variables (e.g., temperature, precipitation, and humidity) and habitat type (including restored habitats), pathogen survival, transmission, and outbreaks. Land-use and land-cover data identify habitat types and areas affected by anthropogenic activities, including restoration, and assess their impact on biodiversity and disease risk.
- **Pathogen Data:** Genomic sequencing data identify specific pathogen strains and track their evolution and spread, offering insights into pathogen dynamics and vulnerable populations. PCR is applied to water, animal, vector, and human samples to detect the presence of specific pathogen DNA or RNA, aiding in the assessment of pathogen prevalence and infection rates. ELISA is used on animal and human samples to detect and quantify antibodies, providing information on exposure.
- **Animal and Vector Data:** Animal data inform on species richness and biodiversity before and after restoration efforts, testing the relationship between biodiversity and disease risk. Vector data provide information on the effectiveness of restoration efforts and interventions for controlling disease vectors.
- **Human Data:** Biological samples and demographic/socioeconomic data collect information to identify populations at higher risk of disease, estimate incidence or cross-sectional prevalence, and evaluate restoration impacts (and biodiversity changes or variations) on human health. Health system data provide information on reported incidence of diseases, the capacity of health systems to respond to outbreaks, and identify surveillance gaps. Perception and behavioral/space use data

assess perceptions of intervention and estimate human exposure to sources of infection in different restored and unrestored habitats.

Data utility, security, and FAIR data

Data Utility

The data produced by BEPREP benefits a range of target groups at local, regional, national, and international levels, including medical doctors, public health agencies, nature conservation organizations, politicians, universities, and research institutes. The data is useful for policymakers, urban planners, landscape designers, pharmaceutical companies, and environmental protection agencies. The project aims to guide future EU-funded nature restorations towards also targeting disease risk mitigation. BEPREP's data is also relevant for ECDC, EFSA, HERA, WHO, IPBES, ETC/BD, EC Knowledge Centre for Biodiversity, forest enterprises (e.g., Stora Enso), and the livestock/agriculture industry.

Data Security

Our consortium ensures that our data security practices comply with relevant regulations and standards, such as the EU's General Data Protection Regulation (GDPR), the ISO/IEC 27001 standard for information security management. BEPREP will have all data stored in REDCap (Research Electronic Data Capture) using University of Helsinki servers. Currently, however, the data is not yet fully integrated within REDCap Helsinki, and part of the data remains with the CSA. We expect full integration by early 2025. Whether for CSA based data (until 2025) or for REDCap especially, the security measures include:

- Encryption: Data in REDCap is encrypted both at rest and in transit using AES-256 encryption and SSL/TLS encryption for secure communication.
- Access Controls: Access to sensitive data is restricted to authorized personnel through user-based, project-based, form-based, and field-level access controls, with user privileges managed by administrators or project owners.
- Regular Backups: We implement automated backups to secure servers, ensuring regular and reliable data backups without manual intervention, verified through REDCap's Data Quality and Compliance module.
- Secure Transfer Protocols: We use HTTPS/SSL encryption, secure file transfer protocols, and SFTP to ensure safe data transfer between different locations while adhering to GDPR requirements, ensuring no data involving EU citizens or generated within the EU is transferred outside the EU.

FAIR Data

To ensure that our data is findable, accessible, interoperable, and reusable (FAIR), we adhere to the following principles and practices:

Making Data Findable, Including Provisions for Metadata

All partners will provide basic metadata records that comply with the OpenAIRE application of the DataCite Metadata Schema to describe the datasets in the BEPREP databases. To ensure findability, interoperability, and re-use, datasets will be described using domain-specific metadata and controlled vocabularies. When necessary, other appropriate standards will be sourced from the Research Data Alliance Metadata Standards Directory.

- **Metadata Standards:** Metadata will follow general standards like Dublin Core, as well as other discipline-specific standards such as the GBIF Metadata Profile, ISO 19115, and INSPIRE metadata standard.
- **Search Keywords:** Metadata will include search keywords to optimize discovery and potential re-use, reflecting important aspects of the data like research question, variables, species, pathogens, and geographic location.
- **Standard Formats:** Metadata will be offered in standard formats like XML or JSON, which can be easily harvested and indexed by search engines and discovery tools.

Making Data Accessible

Data created by partner institutions will be accessible according to their open access/science policies. If datasets are used for commercial purposes, partner institutions will sign agreements with third parties based on fair and reasonable conditions. Access modalities will follow Horizon Europe's recommendations related to open science.

- **Repositories:** All data will be published in repositories compliant with Horizon Europe's open access policy. Initially, data deposited in the BEPREP database will be available upon request to the partner in charge. Relevant datasets will be shared by deposition into open access and trustworthy data repositories with an open license such as Creative Commons CC-BY or equivalent. The European repository, Zenodo, will be the preferred repository, but other domain-specific repositories may also be used.
- **Persistent Identifiers:** Data will always be deposited with a Persistent Identifier (PID) and linked to publications (DOI), ORCID, or researchID and the Research Organisation Registry (ROR) of the institute.

Making Data Interoperable

To ensure interoperability, BEPREP will follow pre-existing standards and vocabularies recommended by scientific communities:

- **Standards and Formats:** We will use Darwin Core for biological diversity, EML for ecological data, ISO 19115 for geolocation, and MxS for sequence data. Data and metadata will be stored and transferred in a comma-separated values (CSV) format or MS Excel-compatible files. Other data formats include .fasta for genetic data, .wav for acoustic data, .sas7bdat (SAS), .RData or .RDS (R), .SAV (SPSS), and .mat (MATLAB). The consortium will reuse existing controlled vocabularies for providing metadata to resources.

Increasing Data Re-use

To facilitate reuse, data will be deposited in relevant repositories under open formats and open licenses such as Creative Commons CC-BY 4.0 or equivalent. Core project partners will be encouraged to openly deposit their data using a Creative Commons BY 4.0 license or equivalent. Data will be stored in repositories such as GENBANK for genetic sequences, GBIF for biodiversity data, and other repositories recommended by the EU registry.

- **Licensing:** BEPREP datasets requiring a license can be shared using the Creative Commons Attribution 4.0 International (CC BY 4.0) or Open Data Commons Attribution License (ODC-By). If the publication of results is not yet finalized, an embargo will be placed on primary data.

- Quality Control and Protocols: For reproducibility of our data analyses, the quality control, bioinformatics, and analytical methods will be made available as annotated protocols on platforms like GitHub, FIGSHARE, or PROTOCOLS.IO in the form of open-source notebooks (e.g., Jupyter and/or R Markdown documents). All outputs will be associated with keywords to increase findability and, once uploaded to GitHub/DRYAD, they will be assigned unique identifiers. Persistent identifiers such as Digital Object Identifiers (DOIs) will be used to ensure that data can be easily located, cited, and linked to over time.

Ethics

The BEPREP project involves the collection of various types of data, including personal data and animal samples, from multiple locations, including non-European countries. The consortium is committed to addressing the ethical requirements associated with the collection, management, and use of such data. All partners have extensive experience in managing these issues, having collaborated on similar projects in the past. Below is a detailed description of the ethical requirements and procedures for BEPREP:

Collection of personal data

- The collection, management, storage, and use of personal data will adhere strictly to the European General Data Protection Regulation (GDPR) principles of lawfulness, fairness, and transparency.
- Personal data will not be shared outside the EU. Within the EU, data sharing will occur only when it benefits the data subjects and appropriate safeguards are in place to protect their rights and freedoms.
- Data subjects will be informed about data sharing activities, and they will receive clear and concise information about their rights regarding the processing of their personal data. This includes rights of access, rectification, erasure, restriction of processing, data portability, and objection to processing.

Collection of Animal Samples

- The collection of animal samples will comply with ethical requirements and procedures for capturing, sampling, and data collection. Necessary authorizations will be obtained to conduct such studies.
- All authorizations will be documented and stored in a secure, accessible location within the consortium's SharePoint, managed by UANT.
- Ethical treatment of animals will be ensured, adhering to humane treatment and welfare standards.

Collection of environmental data

- Data collection in natural environments, particularly in non-European countries, will follow local and national guidelines and legislation. Approval from local ethics committees and/or competent authorities will be sought where necessary.
- Specific measures will be taken to ensure that the data collection does not disrupt local ecosystems or violate any environmental regulations.

Exchange of biological samples between EU and non-EU countries

- The exchange of biological samples will be conducted in strict compliance with EU Regulation (EU) N°11/2014 and Commission Implementing Regulation (EU) N°2015/1866, which relate to the Nagoya Protocol on Access and Benefit Sharing (ABS).
- This ensures that the transfer of biological samples is lawful, benefits the source countries, and complies with international standards for genetic resource sharing.

Ethical management and oversight

- BEPREP has a dedicated work package (WP7) focused on ethical considerations, with the project coordinator holding overall responsibility for the project's ethical management.
- Nivison Junior da Rocha has been appointed as the lead data manager. With over 14 years of experience in data management for longitudinal epidemiological projects in low-income communities, Nivison will oversee adherence to the Data Management Plan by all consortium members and partners, particularly concerning data security and access. The consortium will also recruit in late 2024 a deputy data manager based at Helsinki University.

Compliance with horizon Europe ethical standards

- All activities, whether within or outside the EU, will comply with Horizon Europe ethical standards (deliverable 7.5). This includes ensuring that all data collection, storage, and sharing activities meet the highest ethical standards.

Deliverables in WP7 on ethics

- 7.1: Inclusion and exclusion criteria for human subjects, along with informed consent procedures and documentation.
- 7.2: Detailed information on biological samples collected, including type, quantity, and storage conditions.
- 7.4: Authorizations for facilities and project activities involving animals, ensuring compliance with ethical standards for animal research.

Monitoring and reporting

- Regular monitoring and reporting mechanisms are in place to ensure ongoing compliance with ethical standards. Any deviations or issues identified will be addressed promptly to maintain the project's ethical integrity.
- Ethical considerations and compliance will be reviewed continuously, and any necessary adjustments will be made to adhere to evolving standards and regulations.

By adhering to these detailed ethical requirements and procedures, BEPREP ensures that all aspects of the project are conducted responsibly, with respect for individuals, animals, and the environment.