**Project Acronym:** [LandWorm]

**Project title:** [Impact of Land use and management on earthWorm communities]

**DATA MANAGEMENT PLAN**

# Data Summary

*[What is the purpose of the data collection/generation and its relation to the objectives of the project? What types and formats of data will the project generate/collect? Will you re-use any existing data and how? What is the origin of the data? What is the expected size of the data? To whom might it be useful ('data utility')?]*

## Purpose of the data collection and its relation to the objectives of the project

The LandWorm project (2023-2025) aims to identify and quantify the current and past effects of land use and management on earthworm communities. To achieve this goal, we will aggregate earthworm community and environmental datasets acquired in the past (from the 1970s to 2020s) on the French metropolitan territory. The datasets come mainly from researchers of the LandWorm consortium; no primary data will be created. Earthworm community and environmental data have been collected through various research activities (national or European projects, citizen science projects, theses, ...) and have been published or not. These datasets are stored and owned by different researchers in different/separate files and formats (excel, calc, comma-separated values, ...), see Table 1.

The earthworm community data are derived from field sampling and generally consist of abundance, biomass and taxonomic determination variables. Environmental data refers to any available information on the state of environmental elements (climate, land use and management, physical-chemical properties of soils, etc.). Environmental data were obtained from field observations and/or from the plot manager and could be quantitative and/or qualitative.

## Dataset collection among LandWorm data providers

Data providers commit to share with FRB-CESAB LandWorm project members, earthworm community data (abundance/biomass/taxonomic identification) under their governance, as well as to provide any explanation necessary for their use in the FRB-CESAB LandWorm project. A set of metadata information must be attached to the delivered data, including name and contact details of the person(s) involved in the data collection, collection dates, information about the geographical localization of the sampling sites, units, methods of collection, definition of columns or abbreviations and process (if there have been any changes to the original dataset) and if the data is already published or not. Data need to be shared at the latest after seven months after the beginning of the FRB-CESA My CoReB LandWorm project, i.e. before 06/07/2023.

Raw datasets will be store in My Core which is an online synchronization and data sharing tool managed by the French National Center for Scientific Research (CNRS) and available to any LandWorm members. [Click here to access the folder](https://mycore.core-cloud.net/index.php/s/DSbH3TgNYSQsZcv)

## Dataset homogenisation, cleaning and implementation

The earthworm community datasets will be homogenized, especially the variables concerning taxonomy. The merging and homogenization of the earthworm community datasets will be done with R software, R codes will be deposited on GitHub repository, with a DOI provided from Zenodo (to a stable version of the code).

The environmental variables available in and between the datasets are very heterogeneous and numerous. The LandWorm project consortium therefore decided to select only those environmental variables that are most relevant to LandWorm's goals. This selection of variables will be done by expert opinion to allow a faster and easier merging of datasets. Depending on the availability and completion rate of the environmental variables within the datasets, external databases may be solicited to complete them. The merging, implementation, homogenization of datasets will be done via the R software (R codes will be deposited GitHub repository, with a DOI provided from Zenodo).

The datasets generated at this stage (intermediate/temporary datasets) will be store and available in My Core. [Click here to access the folder](https://mycore.core-cloud.net/index.php/s/DSbH3TgNYSQsZcv)

## Final dataset

Earthworm community and environmental datasets will likely be merged together for output, creating a consistent structure between them using R software (R codes will be deposited GitHub repository with a DOI provided from Zenodo). Each data point will relate to an earthworm community and environmental properties of a plot at a given time. The final data output (as well intermediate steps) will be stored online in CSV format (as a file format that can be inputted into a wide variety of software programs). The size of the data outputs is expected to reach 10.000 data points.

## Data utility outside the project

Outside of the LandWorm project, the data aggregated will be most useful for scientists interested in investigating the impacts of one or more global change drivers on biodiversity. Given the focus is on soil biodiversity, the data should definitely appeal to soil ecologists, however could also supplement other data that above-ground ecologists use as well.

Table 1: Estimation of available data according to the land use studied and the owners, non-exhaustive table

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset name** | **Land use** | **Owners** | **Nomber of plot** |
| SBT-ENI | Agricultural | University of Rennes 1 (D. Cluzeau) | 5000 |
| SOERE-ACBB-Lusignan | Agricultural | University of Rennes 1 (D. Cluzeau) & Institut Agro (G. Pérès) | 132 |
| SOERE-ACBB-Mons | Agricultural | University of Rennes 1 (D. Cluzeau) | 78 |
| SOERE-ACBB-Theix | Agricultural | University of Rennes 1 (D. Cluzeau) & Institut Agro (G. Pérès) | 60 |
| SOERE-PRO-EFELE | Agricultural | University of Rennes 1 (D. Cluzeau) & Institut Agro (G. Pérès) | 50 |
| SoilServ | Agricultural | Institut Agro (G. Pérès) | 92 |
| SUSTAIN | Agricultural | Institut Agro (G. Pérès) | 16 |
| BIO2 | Agricultural | University of Rennes 1 (D. Cluzeau) & Institut Agro (G. Pérès) | 55 |
| RMQS Biodiv | Agricultural and Natural | University of Rennes 1 (D. Cluzeau) & Institut Agro (G. Pérès) | 109 |
| Agrinnov | Agricultural | University of Rennes 1 (D. Cluzeau) | 200 |
| JASSUR | Urban | University of Rennes 1 (D. Cluzeau) | 80 |
| EcobioSoil | Agricultural, urban and natural | University of Rennes 1 (D. Cluzeau) | 1850 |
| Decaëns | Natural | University of Montpellier (T. Decaëns) | 200 |
| La Cage, Feucherolles, Rescape, PSDR Dynamiques, PCRD ANSES | Agricultural | INRAe Avignon (C. Pelosi) | 1000 |
| Bouché 1972 | Agricultural and natural | BETSI database | 1200 |
| #Vers2022 | Agricultural and natural | INRAe Montpellier (M. Hedde) | 250 |

# FAIR data

## Making data findable, including provisions for metadata

*[Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)? What naming conventions do you follow? Will search keywords be provided that optimize possibilities for re-use? Do you provide clear version numbers? What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.]*

At the end of the LandWorm project, all earthworm community datasets that can be made available will be on Dryad. This is a reliable repository that many researchers in our field use and has extensive documentation to facilitate the process. A DOI will be assigned to each dataset, provided by Dryad. In order to aid in data discovery, metada for each dataset will be produced. Dryad provides examples and recommendations for meta-data, so these will be followed (<https://datadryad.org/stash/faq#metadata>). Our will ensure that the metadata includes appropriate keywords in order to aid in the discovery of the data, as well as being able to be harvested and indexed.

## Making data openly accessible

*[Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.*

*Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.*

*How will the data be made accessible (e.g. by deposition in a repository)? What methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)? Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible. Have you explored appropriate arrangements with the identified repository? If there are restrictions on use, how will access be provided? Is there a need for a data access committee? Are there well described conditions for access (i.e. a machine readable license)? How will the identity of the person accessing the data be ascertained?*

*Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible. ]*

In work package 2, we plan to make as many of the earthworm community datasets available to all, but since we are using data provided by other researchers, we will need to ensure that they agree that the data should be made available to all. We propose that at a minimum, total abundance, total biomass, and total richness variables at the plot level be available to everyone. The data will be made available through Dryad, which is free to use and has easy access protocol.

At least at the end of work packages 3 and 4 (i.e. when manuscripts are submitted to a journal), the associated earthworm community datasets will be submitted to Dryad. Upon upload, the datasets will be assigned a unique DOI for each dataset. The data will be made available as soon as the article is published, with no embargo. There will be no legal restrictions on the data, however, specific citations will be provided for researchers to cite if re-using the data.

The data that is being collected for the WPs is not personal or sensitive data, so there is no need for a data access committee. Related to this, the identity of any person accessing the data is not needed to be ascertained.

Environmental data will be made openly available, and licenced under a public domain dedication (CC0). Environmental data will be created towards the start of the work package and then updated upon completion of the work package. This will allow the information on the data can be accessed when the datasets are made available. The data, and associated environmental data, will remain available on Dryad for as long as Dryad is available.

## Making data interoperable

*Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)? What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable? Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability? In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?*

The meta-data format will follow examples provided by Dryad (<https://datadryad.org/stash/faq#metadata>) as well as the Darwin Core (DwC) and the Ecological Metadata Language, which are all common standards inour field. To aid in interoperability and data exchange, our data structure will follow that of the Darwin Core (<https://dwc.tdwg.org/>) whenever possible.

To ensure that any aspect of our datasets that is non-standard are still understandable, in addition to the metadata, we will provide extensive README files that detail all aspects of the datasets, including all ontologies and vocabularies used and how they map to existing ontologies and vocabularies.

As none of our datasets will be primary data, there will be references to other articles or published datasets throughout. For each dataset from each WP, we will provide all relevant and needed references to allow the data to be traced back to it’s original publication or data collector.

## Increase data re-use (through clarifying licences)

*How will the data be licensed to permit the widest re-use possible?*

*When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.*

*Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.*

*How long is it intended that the data remains re-usable?*

Are data quality assurance processes described?

Further to the FAIR principles, DMPs should also address:

We will provide a number of additional documents that will add with data analysis validation as well as facilitating data re-use. We will provide extensive README files with all the datasets, explaining all aspects of the structure and information held within, for example, clarification on any variable units or quality of the data. All code will be provided in publicly available GitHub repositories (with DOIs) to show how all data cleaning, analyses and figures were produced for each WP. These repositories will also contain README files to aid in re-using the code.

We aim to have all data freely available in the public domain at the end of each work package. We will ensure that the provenance of the datasets are thoroughly documented. This will be done by retaining the references of where the data points came from within each dataset, as well as using code for data cleaning and processing, to ensure that any changes to the data are also tracked. Raw data files will also be maintained (as well as intermediary steps) so that changes to the datasets can be “reversed”.

Third parties will also be able to use the data after the end of the project as well. As the data comes from different researchers it will depend on their willingness to share their data in the standards used.

# Allocation of resources

What are the costs for making data FAIR in your project?

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Who will be responsible for data management in your project?

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Dryad's data publishing charges per data submission is $150 US, which is relatively low. The direct cost of data storage/archive will be covered by the LandWorm project. Archiving a pre-print version of the manuscript is free via BioRxiv. GitHub (for code storage) and use of Zenodo for DOIs is also free. As there are no additional costs once everything is online, all of these items will be preserved for as long as the websites/servers are accessible. The post doc of the LandWorm project will be responsible for all data management.

# Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Is the data safely stored in certified repositories for long term preservation and curation?

None of the data generated in this project are considered sensitive data in the traditional sense, so data security regulations are deemed not necessary. For long-term preservation, the data will be stored in Dryad at the end of each work package. In addition, throughout the project, data will be backed up to external hard-drives, institutional servers and cloud storage at regular intervals, in case of hardware failure.

# Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

There will be no ethical or legal issues that can have an impact on data sharing. As there is no personal data included in the data being used in this project, no informed consent is needed for data sharing and long-term preservation.

# Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

There are no other procedures for data management in place (e.g., at the national, sectorial or departmental level).

# Further support in developing your DMP

The Research Data Alliance provides a [Metadata Standards Directory](http://rd-alliance.github.io/metadata-directory/) that can be searched for discipline-specific standards and associated tools.

The [EUDAT B2SHARE](https://b2share.eudat.eu) tool includes a built-in license wizard that facilitates the selection of an adequate license for research data.

Useful listings of repositories include:

[Registry of Research Data Repositories](http://www.re3data.org)

Some repositories like [Zenodo](https://zenodo.org/), an OpenAIRE and CERN collaboration), allow researchers to deposit both publications and data, while providing tools to link them.

Other useful tools include [DMP online](https://dmponline.dcc.ac.uk) and platforms for making individual scientific observations available such as [ScienceMatters](https://www.sciencematters.io).

**SUMMARY TABLE 1**

**FAIR Data Management at a glance: issues to cover in your Horizon 2020 DMP**

This table provides a summary of the Data Management Plan (DMP) issues to be addressed, as outlined above.

|  |  |
| --- | --- |
| **DMP component** | **Issues to be addressed** |
| **1. Data summary** | * State the purpose of the data collection/generation * Explain the relation to the objectives of the project * Specify the types and formats of data generated/collected * Specify if existing data is being re-used (if any) * Specify the origin of the data * State the expected size of the data (if known) * Outline the data utility: to whom will it be useful |
| **2. FAIR Data**  2.1. Making data findable, including provisions for metadata | * Outline the discoverability of data (metadata provision) * Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers? * Outline naming conventions used * Outline the approach towards search keyword * Outline the approach for clear versioning * Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how |

|  |  |
| --- | --- |
| 2.2 Making data openly accessible | * Specify which data will be made openly available? If some data is kept closed provide rationale for doing so * Specify how the data will be made available * Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)? * Specify where the data and associated metadata, documentation and code are deposited * Specify how access will be provided in case there are any restrictions |
| 2.3. Making data interoperable | * Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability. * Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies? |
| 2.4. Increase data re-use (through clarifying licences) | * Specify how the data will be licenced to permit the widest reuse possible * Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed * Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why * Describe data quality assurance processes * Specify the length of time for which the data will remain re-usable |
| **3. Allocation of resources** | * Estimate the costs for making your data FAIR. Describe how you intend to cover these costs * Clearly identify responsibilities for data management in your project * Describe costs and potential value of long term preservation |
| **4. Data security** | * Address data recovery as well as secure storage and transfer of sensitive data |
| **5. Ethical aspects** | * To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former |
| **6. Other** | * Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any) |

|  |  |  |
| --- | --- | --- |
| **HISTORY OF CHANGES** | | |
| **Version** | **Publication date** | **Change** |
| 1.0 | 13.10.2016 | * Initial version |