BIG_PICTURE / Biodiversa+

DATA MANAGEMENT PLAN v0.1

BIG_PICTURE: Developing data management and analytical tools to integrate and advance professional and citizen science camera-trapping initiatives across Europe

Funding Acknowledgement:

This research was funded by **Biodiversa**+, the European Biodiversity Partnership, in the context of the **Big_Picture project** under the 2022-2023 **BiodivMon** joint call. It was co-funded by the **European Commission (GA No. 101052342)** and the following funding organisations: Research Council of Norway, German Federal Ministry of Education and Research, Belgium Science Policy, French Agence Nationale de la Recherche, Italian Ministry of Universities and Research, Netherlands Organisation for Scientific Research, Polish National Science Centre, Ministry of Higher Education, Science and Innovation of the Republic of Slovenia, Swedish National Space Agency, Swedish Environmental Protection Agency and the Spanish Agencia Estatal de Investigación.

Data Management Plan version history

Version	Date	Modification	Author(s)
0.1	28/07/2024	Original draft	Jakub W. Bubnicki Peter Desmet Magali Frauendorf Fabiola Iannarilli John Linnell

Scope of the Data Management Plan

The Data Management Plan (DMP) supports common understanding and expectations among partners and associated partners for the management of data associated with BIG_PICTURE project and serves as a tool to ensure BIG_PICTURE meets research community, partner institution, national and EU standards for data access following Open Science and FAIR principles.

The DMP, as a living document, will be updated throughout the project, and new versions will be made available with each relevant milestone and deliverable. These updates will address the diversity of technological, national, institutional, and individual constraints BIG_PICTURE is aiming to overcome regarding sharing camera trap data at a continental scale.

NOTE
For example, the next version of the DMP will include the legal findings relevant to the European context obtained from WP1, as well as stakeholder considerations from WPs 2 and 3. A dedicated WP1 will investigate legal aspects associated with camera trap deployment in the field and resulting data sharing and storage issues, such as sensitive species, intellectual property rights, and privacy (e.g. images of humans), and provide recommendations and procedures on how to handle these.

Purpose of data collection and sharing

Data collection and research are in accordance with the BIG_PICTURE Project Proposal. The goals of the Project are available online:

- Project website: https://wildlifecamera.eu
- Biodiversa+ website: https://www.biodiversa.eu/2024/04/15/big_picture/

The main objective of the BIG_PICTURE project is to bring together the enormous amount of species data that are collected by thousands of wildlife camera traps (automatic cameras) distributed across Europe by professional researchers, organised groups of citizen scientists, and other private individuals. By providing and developing the appropriate e-infrastructure (data management systems and AI-driven image processing capabilities), enabling frictionless data sharing workflows, and statistical tools for data harmonisation and analysis, the BIG_PICTURE project will facilitate the sharing, integration, and joint analysis of data collected by many different institutions, allowing continental-scale assessments of species' status.

Data collection and documentation

How and what data will be collected, observed, generated, or reused?

The concept of the BIG_PICTURE project is to increase interoperability and facilitate the sharing of existing camera trap datasets between consortium members, associated partners, and external infrastructures (e.g., GBIF, EuropaBON, European Observatory of Wildlife, other Species Information Services). Thus, within the scope of this project, there is no plan to collect new data but to **reuse existing camera trap (CT) datasets**, which we refer to as **Primary data** (referred to as **"background"** in the consortia agreement). The existing CT datasets are defined as those already managed by the consortium and associated partners, as well as those that will be collected and/or processed during the project independently by each partner, such as a result of ongoing research activities or monitoring programs. The BIG_PICTURE activities will also **generate numerous digital outputs** (e.g., software tools, AI algorithms, statistical approaches, reference image libraries, and analysis routines) that we describe below in the **Secondary data** section.

Primary data

The primary data consists of **CT datasets** collected and/or managed by members of this Consortium and associated partners. These datasets contain 1) **CT media** (images and videos) and 2) **CT tabular data** (metadata and observations) and can range from thousands to millions of observations each. All of the datasets that will be included in this project will be standardised and harmonised using the **Camera Trap Data Package (Camtrap DP)** format which is described in detail in the next section.

The **CT media** files are images (e.g. JPG) and videos (e.g. MP4) captured by camera traps. These data, together with associated metadata, will be used to build a shared, high-quality library of annotated media of European wildlife species for the purpose of **developing and validating AI-based methods for automatic processing of CT data**. The library will allow researchers and developers to train new, more robust AI models for the classification of CT

footage at the level of objects, species, sex, individuals, behaviour, as well as experiment with new AI architectures published in the literature (e.g., Multimodal AI Models). An important aspect will also be the ability to apply AI to extract **background information** from CT media, such as weather conditions, snow depth, or vegetation type. After careful expert-based validation, a subset of CT media from all available wildlife observations will be used as **a gold-standard dataset to build a shared training and validation library**. The selection will be done in a way that ensures the representativeness of the library for European wildlife species and habitats to the maximum extent possible.

The **CT tabular data** are metadata and observations of wildlife, human activity and environment, associated with the media files. Metadata includes information about the project and camera trap sampling scheme and protocol, such as e.g. the location, date and time of the recording, the camera settings, as well as other technical details on the camera trap deployment and media format and location (e.g. in a cloud storage). Observations are the results of the expert-based and/or AI-assisted classification of the visual content of the raw media files into scientific information describing the presence, biological features (e.g. sex, age, behaviour) and number (group size) of wildlife species and background information (e.g. weather conditions, vegetation type, human activity). Following the **Camtrap DP** standard, the tabular data will be exchanged in a structured **CSV** format and linked to the media files using unique identifiers (see the next section).

NOTE
At this stage of the project, we are not able to provide a detailed list of all primary **CT datasets** that will be used in this project, as the Consortium is still in the process of collecting information on available resources through **an online survey and interviews with partners**. However, as this project is built on a large consortium of 17 institutions based in 10 countries with access to data from >100 sites in 30 European countries (with additional data from 20 more countries outside of Europe if needed) we expect, based on initial inventory, tens of millions of images and videos of wildlife species and background information to be compiled into joint datasets. This section will be updated with the results of the ongoing survey in a next version of the DMP.

Another primary data that will be used in this project is **satellite remote sensing data**. We will use the environmental information contained in the background of the CT media, linking it to satellite-derived remote sensing data to explore novel ways to link satellite and terrestrial remote sensing. Remote sensing data from satellites is an important, continent-wide, data source that we will use in WPs 8-10 to link wildlife occurrences to habitat characteristics (e.g., to allow extrapolation to larger areas).

Secondary data (digital outputs)

The major **Secondary data** products will include:

1. **Integrated and harmonised CT datasets** derived from the **Primary data** already collected and managed by the members of this consortium and all associated partners. These datasets will be used in a series of **case studies** that demonstrate the utility of our toolkits and the added value of pan-European data sharing (WPs 10 and 11).

2. **AI models** for automatic processing of wildlife observations and extraction of background information from CT media. These models will be trained (and/or fine-tuned) and validated using the subset of **Primary data** compiled into a gold-standard dataset to build a shared library.

3. **Statistical approaches & data analysis routines**; data analysis pipelines, R & Python scripts, Jupyter Notebooks. These routines will be developed and validated using the **Primary data** and **Integrated and harmonised CT datasets**.

4. **Software tools** for data management, analysis, and visualisation. These include AI training pipelines, R packages, Python libraries and/or web applications.

Other data

Personal data of members, associated partners, and stakeholders will be collected and processed in the context of the project. These data will be used for data collection, communication, project management, and reporting purposes. The data will be stored in the project's **internal secure databases** and will be shared only within the consortium members group. The data will be processed in accordance with the **General Data Protection Regulation (GDPR)**. Names, professional titles, phone numbers, and email addresses of stakeholders will be collected. These data are collected for the sole purposes of stakeholder engagement (two-way communication - data-gathering and providing feedback) and communication as defined in the project plan on the legal basis of public interest. Stakeholders may have the right to have their information erased, corrected, or limited. Their rights can be exercised by contacting the project coordinator. This list will not be shared or made public without the consent of the stakeholders.

What documentation and metadata will you provide with the data?

The exchange, integration, and harmonisation of **CT datasets**, both internally during the life of the project and after the project completion, will be based on a published standard for the exchange of camera trapping data, **Camtrap DP**. This standard is actively developed and maintained by members of this consortium. As Camtrap DP is directly derived from the Data Package specification, it automatically inherits most of the basic principles of **FAIRness**. Camtrap DP is now maintained under the umbrella of the Biodiversity Information Standards (TDWG), and GBIF supports it as one of their data publication formats. Using Frictionless Data Package specification, a Camtrap DP dataset contains two types of files: **CSV files with tabular data** and a **JSON descriptor file with dataset-level metadata**. A comprehensive description of Camtrap DP, together with its technical documentation and a guide on best practices for managing and publishing camera trap data in general, has already been published by members of the Consortium and is available online:

- Camtrap DP website: https://camtrap-dp.tdwg.org/
- Bubnicki JW, Norton B, Baskauf SJ, Bruce T, Cagnacci F, Casaer J, Churski M, Cromsigt JPGM, Farra SD, Fiderer C, Forrester TD, Hendry H, Heurich M, Hofmeester TR,

Jansen PA, Kays R, Kuijper DPJ, Liefting Y, Linnell JDC, Luskin MS, Mann C, Milotic T, Newman P, Niedballa J, Oldoni D, Ossi F, Robertson T, Rovero F, Rowcliffe M, Seidenari L, Stachowicz I, Stowell D, Tobler MW, Wieczorek J, Zimmermann F, Desmet P (2023). Camtrap DP: an open standard for the FAIR exchange and archiving of camera trap data. Remote Sensing in Ecology and Conservation. https://doi.org/10.1002/rse2.374

- Reyserhove L, Norton B & Desmet P (2023) Best Practices for Managing and Publishing Camera Trap Data. GBIF Secretariat: Copenhagen. https://doi.org/10.35035/doc-0qzp-2x37

The documentation for the **Integrated and harmonised CT datasets** will be provided alongside the datasets in research repositories (e.g., Zenodo or Dryad). The documentation will include a detailed description of the data sources, data processing steps, and data harmonisation procedures. The datasets will be accompanied by **metadata** that provides a brief overview of the dataset, its structure, and the data fields. The metadata will also include information on how to access and use the dataset, as well as any specific requirements or limitations associated with the data. More detailed descriptions of the analytical steps will be provided in the form of **scientific publications** (including data papers) and/or **technical reports** published, whenever possible, in a gold open access mode and licenced with a Creative Commons Attribution licence (CC BY).

The documentation for developed **AI models**, **Statistical approaches & data analysis routines**, and **Software tools** will be provided in the form of **technical documents, source code comments, tutorials (e.g. Jupyter Notebooks or R Markdown), and user manuals**. The technical documentation will provide a detailed description of the code structure, algorithms, and parameters used in the development of the models, approaches, and tools. The tutorials will include step-by-step instructions on how to use the code, as well as examples of input and output data. The user manuals will provide a more general overview of the models, approaches, and tools, as well as instructions on how to install and run them. The documentation will be made openly available on GitHub or similar platforms, alongside the source code, preferably under the MIT, GNU GPL or CeCILL licence.

Data sharing and reuse

How and where will the data be shared?

Primary data

The **Camtrap DP** data exchange standard is designed to be machine and human readable and to permit **Findability**. Adopting this data standard and ensuring **Interoperability** in practice is one of the key objectives of this project. This includes the adoption of the data standard by CT data management platforms (already implemented in Agouti and TRAPPER, maintained by consortium members) and data analysis tools. The aim is to facilitate an interoperable and robust data flow between all relevant CT infrastructure components and participants in Europe and beyond. Additionally, the format will make it easier to publish selected data to external e-infrastructures such as GBIF, EuropaBON (and the emerging European Biodiversity Observation Coordination Centre, EBOCC) or the European Observatory of Wildlife. Data from partners without sufficiently developed infrastructure will be hosted for them on either the TRAPPER or Agouti platform. The **Primary data** will be potentially **Accessible** within the participating institutions existing databases (they already maintain databases with millions of records), although access rules will need to be clarified as the project progresses following the outcomes of WPs 1-4 which will influence the **Reusability** of the raw data. In this way we will strive to be as open as possible, and as closed as necessary with the **Primary data**, respecting national legislation and institutional policy and accommodating requirements for sharing data on species of conservation concern.

A dedicated **CT Metadata Hub** will be the main interface (**UI & API**) for sharing information about the available **Primary data** in **Camtrap DP** format within the project and with external partners. The CT Metadata Hub will be designed to allow for the controlled sharing of metadata and data, with the possibility of setting up access rules for different users. It will build on the data sharing experience of the existing communities and stakeholders'

networks managed by partners of the consortium (EUROMAMMALS, SCANDCAM, and ENETWILD) and will be designed to be interoperable with major CT data management platforms operating in Europe (e.g., Agouti, TRAPPER, SCANDCAM, Wildlife Insights, Camelot). The CT Metadata Hub will be implemented in WP5 (Infrastructure) and will provide a user-friendly interface and API for external access to the data. It will allow for real-time searches for specific datasets and extraction of data from the integrated Europe-scale camera trap datasets. The hub will be based on available **fully-featured open-source software for (meta)data catalogues** such as CKAN, Dataverse, or Zenodo. Alternatively, a more lightweight and tailored solution to BIG_PICTURE needs will be developed. The CT Metadata Hub will be designed to be easily scalable and maintainable and will be hosted on the infrastructure of one of the Project members.

An **internal data model** will be developed to facilitate **the integration of CT tabular data** in **Camtrap DP** format from different sources. This means harvesting data packages and storing combined data in one database, ensuring that the data can be easily shared and reused via the **CT Metadata Hub** interfaces. A draft of such a solution is currently under development by the **EuroCam** initiative (part of the EUROMAMMALS network) and will be tested in this project.

Moreover, access to integrated Europe-scale camera trap datasets will be facilitated by dedicated **high-level software tools** (R and Python packages, Jupyter Notebooks) to read and analyse data exported in **Camtrap DP** format. For example, the camtrapdp R package, developed by members of this consortium, is designed to make working with Camtrap DP as frictionless as possible. It can be used as a standalone package or as a dependency for other R packages like camtrapR and camtraptor. Currently, it offers functionality for reading, filtering, and manipulating data, as well as translating data to Darwin Core Archive and EML (used by GBIF). We plan to extend it with functionality for merging and writing datasets. This R package can also be used as a dependency for a Shiny dashboard to create human-readable reports and interactive data exploration dashboards.

Secondary data

1. **Integrated and harmonised CT dataset** needed for analysis replication, along with other products of the data processing and analysis pipelines (e.g., compiled Essential Biodiversity Variables), will be deposited as FAIR and open data in research repositories (e.g., Zenodo or Dryad), subject to the specificities of WP1 and specific national and institutional considerations of the individual dataset owners.

2. **AI models**. Parameters (weights) of trained and/or fine-tuned AI models will be made openly available on dedicated repositories such as GitHub or HuggingFace, preferably under licences such as MIT, GNU GPL, OpenRAIL, Creative Commons, or other licences depending on the product.

3. **Statistical approaches & data analysis routines** (data analysis pipelines, R & Python scripts, Jupyter Notebooks) will be made openly available on GitHub, preferably under the MIT, GNU GPL licence.

4. **Software tools** for data management, analysis, and visualisation. These include AI training pipelines, R packages, Python libraries and/or web applications. The source code of software tools developed in this project will be published on GitHub or similar platforms, preferably under the MIT, GNU GPL or CeCILL licence. We will seek peer review (e.g., from rOpenSci) to ensure quality, ease of use, and long-term viability.

5. **Scientific publications & technical reports**. Whenever possible, we will choose gold open access and a Creative Commons Attribution licence (CC BY) for scientific publications written as part of this project. Publications that we can self-publish (e.g., DMP) will be licensed under CC BY. All partners have set aside resources in their budgets for article processing costs.

Criteria for external pre-publication sharing

NOTE

To be developed in the next version.

Are there any necessary limitations to protect sensitive data?

The observations of sensitive species (e.g. critically endangered species) and all associated metadata will be excluded or post-processed (e.g. coordinates generalisation, image background subtraction, etc.) before data publication. Sensitive data in respect to privacy, such as recordings of humans or vehicles will be automatically anonymized or deleted before publication following the EU General Data Protection Regulation (GDPR).

NOTE
This section will be updated with the results of the WP1 that will investigate typical camera trap data sharing issues, such as sensitive species, intellectual property rights, and privacy (e.g. images of humans), and provide recommendations and procedures on how to handle these.

Ethics, legal and security issues

How will ethical issues be addressed and handled?

NOTE
This section will be updated with the results of the WP1 that will investigate typical camera trap data sharing issues, such as sensitive species, intellectual property rights,

and privacy (e.g. images of humans), and provide recommendations and procedures on how to handle these.

How will data access and security be managed?

Each institution contributing data is bound by different rules concerning access, related to funding bodies, national privacy rules, rules concerning release of sensitive information on endangered species and the agreements made between researchers and citizen scientists that contribute information. A set of WPs (WP 1, 2, 3, 4) is designed to clarify legal, institutional and individual concerns about sharing data openly with the explicit goal of making as much of the underlying data as openly available as possible or clarifying specific procedures for making it open.

All participating institutions already maintain strong security on their e-infrastructure, and by building on existing structures there is no need to create a new system. However, it is clear that linking systems and enhancing interoperability will require a rethink of access procedures, where solutions will vary between institutions depending on their underlying data architecture.

How will you handle copyright and Intellectual Property Rights issues?

NOTE
This section will be updated with the results of the WP1 that will investigate typical camera trap data sharing issues, such as sensitive species, intellectual property rights, and privacy (e.g. images of humans), and provide recommendations and procedures on how to handle these.

Data storage, preservation and costs

How will your data be stored and backed-up during the research?

By the data curators of existing platforms and databases within the participating partners - using systems such as TRAPPER, Agouti, and SCANDCAM, as well as their in-house, national, or commercial storage infrastructures. The costs of data archival are already being maintained by the participating partners. The BIG_PICTURE project will only cover costs associated with enhancing the interoperability of systems or paying the one-time costs for processed data archival necessary for replication of published analyses and keeping our digital tools available online beyond the project's lifespan.

What is your data preservation plan?

Primary data, including both **CT media** and **CT tabular data**, will remain in the various institutional databases, which represent a sustainable archive that will continue to grow as the contributing projects continue. Any processed data, i.e., **Integrated and harmonised CT datasets**, **AI models**, **Statistical approaches & data analysis routines**, necessary for replication of analyses associated with scientific publications, will be deposited in research repositories (e.g., Zenodo or Dryad) and made openly available. The **Software tools** produced by the project will be made openly available on GitHub or similar platforms. A large library of **CT media** (a subset that can be shared openly) will be made available on the internet for future AI developers. See the section How and where will the data be shared? for more details.