MAMBO

MODERN APPROACHES TO THE MONITORING OF BIODIVERSITY

D2.2 Guidelines for data standards and technical specifications of MAMBO tools

08/11/2023

Lead beneficiary: INRIA

Author/s:

Funded by the European Union

D2.2 Guidelines for data standards and technical specifications of MAMBO tools

Prepared under contract from the European Commission
Grant agreement No. 101060639
EU Horizon Europe Research and Innovation Action

Project acronym:	**MAMBO**
Project full title:	**Modern Approaches to the Monitoring of Biodiversity**
Project duration:	01.09.2022 – 31.08.2026 (48 months)
Project coordinator:	Dr. Toke Thomas Høye, Aarhus University (AU)
Call:	HORIZON-CL6-2021-BIODIV-01
Deliverable title:	
Deliverable n°:	D2.2
WP responsible:	Dr. Niels Raes
Nature of the deliverable:	Report
Dissemination level:	Public
Lead beneficiary:	INRIA
Due date of deliverable:	M14
Actual submission date:	M14

Deliverable status:

Version	Status	Date	Author(s)
1.0	Toc	01/06/2023	Alexis Joly (Inria)
2.0	Draft	21/07/2023	Alexis Joly (Inria) Niels Raes (Naturalis)
3.0	1st part completed	24/08/2023	Alexis Joly (Inria) Niels Raes (Naturalis) Dan Stowell (Naturalis)
4.0	all parts completed	13/10/2023	Alexis Joly (Inria) Niels Raes (Naturalis)
5.0	Internal review and last revisions completed	05/12/2023	Pierre Bonnet Toke Thomas Høye Alexis Joly

Table of Contents

Summary

This deliverable provides guidelines about the data standards and technical specifications to be used by MAMBO tools so as to make them FAIR (Findable Accessible Interoperable and Reusable) at the end of the project. Rather than developing a n-th new system in MAMBO, our proposal is to rely primarily on existing large-scale open data infrastructures. Firstly, in section 1, we describe all the platforms planned or recommended for managing and sharing the biodiversity data produced within the project, and/or for hosting the tools developed. In section 2, we describe the recommended standards in terms of data storage and exchange formats. Finally, section 3 presents detailed guidelines/recommendations for each of the innovative technologies developed within MAMBO.

1 Recommended data platforms for the management and sharing of biodiversity data

Relying on large-scale open data infrastructures offers several advantages over developing a new system for storing data: (i) it eliminates the need for major investments in hardware, software and maintenance, thus reducing operating costs, (ii) open data infrastructures are typically designed to handle large volumes of data and can easily scale to accommodate growing data needs, ensuring long-term sustainability, (iii) these infrastructures often follow industry standards and best practices, promoting data interoperability and making it easier to integrate with other systems and collaborate with external partners, (iv) open data initiatives often have a community of users and contributors, providing valuable resources, support, and a wealth of shared knowledge, (v) established open data platforms often have robust security measures and redundancy in place, enhancing data protection and ensuring data availability, (vi) utilising open data infrastructures grants access to a wider ecosystem of data sources, tools, and analytics, fostering innovation and data-driven decision-making.

We list hereafter the biodiversity data platforms recommended for the management and sharing of biodiversity data collected or produced within MAMBO.

GBIF (for species occurrences and survey data): GBIF, the Global Biodiversity Information Facility, is an international network and research infrastructure that aims to provide free and open access to biodiversity data from around the world. GBIF was established in 2001 and is governed by

an intergovernmental agreement. It implements data quality control measures to ensure the reliability and accuracy of the shared data. The data available through GBIF is governed by Creative Commons licences (CC0, CC-BY and CC-BY-NC). To date, GBIF includes more than 89,000 datasets on its portal published by more than 2,100 data publishing institutions including several MAMBO partners: Pl@ntNet (CIRAD and Inria), Naturalis.

The GBIF platform is specifically designed for sharing biodiversity occurrence data, i.e. observations of species in specific locations and time periods. These occurrence records are typically based on individual observations of species in the wild. While GBIF was not initially optimised for sharing biodiversity survey data, researchers can now contribute relevant data to GBIF if they convert their survey data into species occurrence records and use the appropriate metadata fields of the DarwinCore standard. Comprehensive guidelines are provided in this regard, in particular about how to link records using an event identifier data element (eventID). In this way, many extension records can exist for each single core event record.

GBIF offers the two main ways to access the published data:

- A Web GUI for general users: the most used application of GBIF is a user-friendly web interface where general users can explore and access biodiversity occurrence data using interactive functionalities. The main key features are:
    - Interactive Maps: Users can visualize species occurrence data on interactive maps, enabling exploration of species distribution patterns across regions and time.
    - Search Functionality: The GUI allows users to perform specific searches based on species names, locations, and other criteria to discover relevant biodiversity data.
    - Occurrence Details: Detailed information about each occurrence record, including taxonomic details, collection date, and locality, can be accessed through the GUI.
    - Taxonomic Hierarchy: The GUI supports exploring species within their taxonomic hierarchy, enabling users to browse data at different taxonomic levels.
    - Species Profiles: Users can access species profiles, providing an overview of species distribution, occurrence data, and related information.
    - Occurrence Download: The GUI allows users to download biodiversity occurrence data for further analysis and research.
    - Data Visualization: Various data visualization tools and charts aid in understanding biodiversity trends and patterns.

- An API for developers: GBIF offers an Application Programming Interface (API) that allows developers to programmatically access and integrate biodiversity occurrence data into their own applications, websites, and analyses. The GBIF API allows users to query the database, retrieve specific data sets, and perform various operations on the data. It supports a wide range of functionalities, including searching for species occurrences, retrieving taxonomic information, obtaining species distribution data, and accessing information about data publishers and datasets. Queries are expressed as HTTP requests to the GBIF API endpoints, specifying parameters such as taxonomic names, geographic regions, time ranges, and data formats to retrieve the desired information. The API responses are typically returned in JSON (JavaScript Object Notation) format, which is a commonly used data interchange format. A comprehensive documentation for the API is provided, including detailed information about the available endpoints, query parameters, response formats, and authentication methods. Developers can refer to the documentation to understand the capabilities of the API and learn how to effectively utilize it in their projects.

Among the many advantages of the GBIF platform is a system for tracking scientific publications based on DOI. This makes it possible to find out about all the scientific publications that have used data from a particular provider.

EBV data portal (for raster datasets related to essential biodiversity variables): The EBV Data Portal is an eShape initiative sharing a variety of raster datasets related to Essential Biodiversity Variables (EBV). It develops a platform for distributing and visualising EBV datasets. In particular, it contains a geographic cataloguing system that supports a large number of spatiotemporal and EBV specific attributes and enables their discoverability. To facilitate user interaction, it offers a web-based interface where users can upload, discover and share essential biodiversity spatiotemporal data through intuitive interaction with cataloguing and visualisation tools. Using the EBV Catalogue, the user can explore the characteristics of the data based on the definition of the EBV Cube standard (which is itself based on the NetCDF data format standard). The Catalogue also allows browsing of the description of the metadata in the specifications of the Attribute Convention for Data Discovery (ACDD) and in the Ecological Metadata Language (EML) vocabulary. This enables easy interoperability with other metadata catalogues.

Pl@ntNet (for image-based plant observations and image-based vegetation plot surveys): Pl@ntNet is a participatory platform allowing users, including botanists, researchers, and nature enthusiasts, to contribute to the identification and study of plants by sharing photos and observations of plant species. It is piloted by a consortium of research organisms including two MAMBO partners (Inria and CIRAD). The data available through Pl@ntNet is governed by the Creative Commons Attribution licence CC-BY-SA. Each observation is assigned a Unique Resource Identifier hosted on https://identify.plantnet.org/. The platform was initially optimised for individual specimen observations but it is currently being extended (in the context of MAMBO and the GUARDEN EU project) to also manage vegetation plot surveys (with multiple specimens associated to the same plot and the possibility to upload multi-specimen images such as cover images, quadrat images, side views, etc.). The platform will also be extended to enable batch import of large quantities of observations through a web GUI or via a web service.

Key features of Pl@ntNet platform include:

- Automated identification: users can receive automated suggestions on the species' identity based on the images of the plant(s) processed by an AI model.
- Cooperative learning: the plant observations can be reviewed by the community and a user can typically confirm the initial species name shared by the observer (with the help of the AI) or propose an alternative species name. Only the observations that reach a sufficient degree of confidence are labelled as valid observations and then added to the training set of the AI model and shared in open platforms.
- Geographic Coverage: Pl@ntNet collects data from users worldwide, providing a global perspective on plant distribution and diversity.
- Research and Conservation: the data gathered through the platform is open and can be used for scientific research, biodiversity monitoring, and conservation initiatives. In particular, validated Pl@ntNet occurrences are shared in GBIF.

Observation.org (for citizen science recordings): Observation.org is a participatory platform allowing users, including citizen scientists, amateur experts and professionals, to contribute to the identification and study of fungi, plants and animals by sharing photos and observations. The website is maintained by Observation International Foundation and has as two main missions : (i) to share observational data about global biodiversity, past and present, as a source of knowledge for the

future, and (ii) to facilitate observers around the world through a multilingual global observation system with a species registry for all known species and species groups in nature, flora and fauna, and share a data collection of validated field data through that system with anyone anywhere in the world. Observation.org and its sibling websites Waarneming.nl and Waarnemingen.be annually collect over 30 million observations from all over Europe which are brought together by over 100.000 active users. Their database contains over 78 million images of plants and animals most of which are validated by amateur-experts naturalists. This database forms the core of the training data used to build image recognition for European species. Annually all observations are uploaded to GBIF. The emphasis of data and users is on northwest Europe but the number of users is strongly growing in other parts of Europe. An increasing number of natural history museums and nature organisations are using Observation.org as their main portal to organise citizen science activities. Sharing multimedia with third parties depends on the licences set by the author. For new users, image and sound licences are set to CC-BY-NC by default. Registered users can change their multimedia licences in the user settings but most keep the settings on CC-BY-NC. Observation.org is in contact with existing National Biodiversity Portals in other countries and has no ambition to compete with them. As part of the work on image recognition for European species Observation.org co-operates with the following partners: Artsdatabanken Norway (NIBC, Norway), SLU Artdatabanken (SSIC, Sweden), UK Centre for Ecology & Hydrology (CEH, UK), Arter (Denmark), Finnish Museum of Natural History Luomus (Laji, Finland), Senckenberg (Germany), Waarneming.nl (Netherlands), Waarnemingen.be – (Belgium). All these partners share training data and split the cost for deployments of AI-models.

Xeno-canto (for validated recordings of bird sounds): There are various websites including Observation.org and iNaturalist where people can upload sound recordings. Here citizen scientists can upload recordings and combine them with metadata resulting in useful observation records. However, the quality of recordings and the intensity of curation of recordings is often limited compared to the quality of images and their curation on these same portals. This makes that the recordings of these portals cannot readily be used as training data for sound recognition. For that reason, an additional portal was needed which is focused on higher quality (both in identification, curation and technical quality) sound recordings. Instead of developing a new portal the EU Horizon projects MAMBO, GUARDEN and TETTRIs opted to use the existing platform Xeno-canto. Xeno-canto is a website for sharing recordings of wildlife sounds from all across the world. Xeno-canto is maintained by a small team of volunteer admins with crucial assistance from Naturalis Biodiversity Center and help from the

xeno-canto community. Xeno-canto is run by the Xeno-canto foundation (or officially Stichting Xeno-canto voor natuurgeluiden), a charity from the Netherlands. Xeno-canto uses the ever evolving possibilities of the internet to: popularise wildlife sound recordings worldwide, improve accessibility of wildlife sounds, & increase knowledge of wildlife sounds and the website is focused on enjoyment, education, conservation, and science. The open access policy of Xeno-canto made the global community of people interested and working on wildlife sound to adopt Xeno-canto as their main portal. This has not only resulted in large numbers of people uploading recordings but also in a large number of specialists helping with the curation of data. At present Xeno-canto with over 700,000 recordings makes for the largest open access database of wildlife recordings. Although the focus was traditionally on birds the website opened up for grasshoppers and bats in the last two years. The benefit for MAMBO to use Xeno-canto as its main source of data is that the data is of high quality and from one source which reduces the time needed for data management. The coverage of data is such that a recognition model for breeding birds of Europe could be built for MAMBO without the use of additional sources. Currently MAMBO and Xeno-canto are cooperating at improving the geographical coverage of data for bats and grasshoppers in Europe which will allow MAMBO to use Xeno-canto data as the main source for the training data of the recognition models for these species groups as well.

GLOBI (for species interaction data): The Global Biotic Interactions network (GloBI) provides open access to finding species interaction data (e.g., predator-prey, pollinator-plant, pathogen-host, parasite-host) by combining existing open datasets using open-source software. It is the only large-scale open data infrastructure storing this type of data across all kingdoms.

ZENODO (for any kind of data): Zenodo is a digital repository and platform for open-access research data and scholarly outputs. It is operated by CERN, the European Organization for Nuclear Research, and supported by the European Commission. Researchers can use Zenodo to deposit and share a wide range of research outputs, including datasets, publications, software, and multimedia content, making their work more accessible to the global research community. Zenodo assigns a DOI (Digital Object Identifier) to each deposited item, ensuring its long-term accessibility and citability. It is a valuable tool for promoting transparency, collaboration, and the open sharing of scientific knowledge across various disciplines.

DRYAD (for any kind of data): Dryad is a digital repository and data preservation platform specifically designed for the storage, sharing, and long-term preservation of research data associated with scientific publications. Researchers can use Dryad to deposit their datasets, making them openly accessible to the scientific community. This service plays a vital role in promoting transparency and reproducibility in scientific research by ensuring that datasets are available for scrutiny and reuse. Dryad is commonly used in various scientific fields to archive and share data, contributing to the broader open science movement. Compared to ZENODO, Dryad primarily focuses on research data associated with scientific publications, especially in the life sciences, environmental sciences, and related fields. It often integrates closely with the publication process, allowing researchers to link datasets directly to their articles. It also allows to share larger datasets (with payment based on volume) whereas ZENODO is limited to 50 Gb per dataset (but free).

EIDC: The Environmental Information Data Centre (EIDC) is part of the Natural Environment Research Council's (NERC) Environmental Data Service and is hosted by the UK Centre for Ecology & Hydrology (UKCEH). It manages nationally-important datasets concerned with the terrestrial and freshwater sciences. It will receive the following data types and formats: Tabular data (Comma-separated = CSV, or tab-delimited = TAB), Spatial raster data (GeoTIFF), Spatial vectorial data (Geopackage, SpatiaLite), Images (PNG, JPEG), Movies (MPEG, MP4, MOV, AVI), Sound (MP3, WAV).

The EIDC requires two types of metadata: discovery metadata and contextual metadata. Discovery metadata is published in EIDC's data catalogue and includes simple information such as:

- A brief, concise title (see guidance)
- A short description of the dataset (guidance)
- Brief information about how the data were created/processed (guidance)
- A list of those who created the dataset (authors)

Discovery metadata also contains information about how to access (download) the dataset, the terms and conditions regarding its use and how others using it should acknowledge & cite the data.

Contextual metadata is requested to help researchers to comprehend and reuse the data appropriately and to help to avoid *mis*use and *mis*understanding of the data. There is no standard way of recording contextual metadata. EIDC policy is to make supporting documentation available with the data as a separate, linked document(s) in the following preferred formats: docx, csv, txt.

Each repository receives a DOI.

EOSC (for web services): The European Open Science Cloud (EOSC) is a European Union initiative aimed at creating a collaborative environment for researchers and scientists to access and share data, tools, and resources across disciplines. It serves as a digital platform to facilitate open and seamless access to research data and services. EOSC promotes transparency and openness in research, fostering innovation and collaboration in the European research community. It plays a pivotal role in advancing the goals of open science and accelerating scientific discoveries through enhanced data sharing and accessibility. Publishing web services or tools in the EOSC catalog provides crucial visibility within the European Open Science Cloud (EOSC) ecosystem, increasing the likelihood of adoption by researchers and institutions engaged in open science practices. It fosters collaboration and interoperability with other EOSC resources, enabling seamless data sharing and integration into research workflows.

Github (for source codes): GitHub is a web-based platform that provides version control and collaborative tools for software development. It allows developers to store and manage their code repositories, track changes, collaborate with others through features like pull requests and issues, and host their open-source projects. GitHub has become a central hub for software development, enabling efficient code management and collaboration among developers worldwide.

2 Recommended standards and data formats for the sharing of biodiversity data & biodiversity-related services

We list hereafter the standards (or common exchange formats) recommended for the development of MAMBO technologies.

Standards recommended for in situ observations:

- DarwinCore: DarwinCore, short for Darwin Core Standard, is a standardised data schema or set of terms specifically designed for biodiversity data. It provides a common framework and vocabulary for describing and sharing information about species occurrences, taxonomy, geographic location, and associated data such as measurements and observations. DarwinCore enables interoperability and facilitates the exchange and integration of biodiversity data across different systems and platforms. The extension Darwin Core "event", in particular, facilitates the report of sampling events and associated taxa or recovered specimens.

- AudioVisual Core: AudioVisual Core (previously "Audubon Core") is a standardised metadata schema that provides a framework for describing and sharing information about biodiversity multimedia resources, such as images and sounds, with a focus on bird-related data and collections.
- Humboldt Core: Humboldt Core is a community-developed standard for representing critical information about scope, method and completeness of biological inventories. It provides a means for standardised capture of information that is typically reported in any inventory. The standard has been developed to be usable across the wide range of inventories that are performed, and has been rigorously tested to assure its usability. Terms in the Humboldt Core have been carefully cross-walked to other biodiversity data standards to assure compatibility where possible with other data dictionaries.
- SensorThings: SensorThings is an OGC standard that facilitates the exchange and management of sensor data within the Internet of Things (IoT), enabling seamless integration and interoperability of sensor observations for diverse applications.
- Camtrap DP: Camera Trap Data Package (or Camtrap DP for short) is a community developed data exchange format for camera trap data. A Camtrap DP is a "Frictionless Data Package".

Standards and common resources used for taxonomic data (species names in particular):

- GBIF backbone: The GBIF backbone, also known as the GBIF Occurrence Backbone, refers to a key component of the Global Biodiversity Information Facility (GBIF) infrastructure. It is a curated and standardized dataset that provides a comprehensive compilation of occurrence records from various sources, including museums, herbaria, research institutions, and citizen science initiatives. The GBIF backbone serves as a reference dataset for biodiversity occurrence data, integrating and harmonizing diverse sources to ensure consistency and interoperability. It forms the foundation for biodiversity research, analysis, and conservation efforts worldwide. The GBIF backbone will soon become the extended Catalogue of Life (COL) as GBIF and COL are working on this together.
- POWO: POWO (Plant of the World Online) is an online database and resource managed by the Royal Botanic Gardens, Kew. It aims to provide comprehensive and up-to-date information on plant names, taxonomy, distribution, and other relevant botanical data. POWO serves as a valuable tool for researchers, botanists, and plant enthusiasts, offering access to a vast

collection of plant species information from around the world, including accepted names, synonyms, common names, distribution maps, and more.

- IPNI: The International Plant Names Index (IPNI) is a comprehensive database that serves as a global index of published plant scientific names, providing authoritative information on plant names, synonyms, and associated taxonomic literature.
- TDWG Authors of Plant Names : TDWG Authors of Plant Names is a database that catalogues and provides information about the authors or author teams responsible for establishing and publishing botanical plant names, aiding in proper citation and recognition of their contributions.

Additional standards used for biodiversity characterization:

- WGSRPD (TDWG): the World Geographical Scheme for Recording Plant Distributions (WGSRPD) is a standardized system developed by the International Working Group on Taxonomic Databases for Plant Sciences (TDWG) to classify and record plant distribution data based on geographical units. This system allows organizations to compare and exchange plant distribution data while avoiding data loss caused by incompatible geographical boundaries. The WGSRPD classifies geographical units into four levels: continental, regional (or subcontinental), botanical country, and basic recording units, considering both botanical and political considerations.
- IUCN status: the IUCN (International Union for Conservation of Nature) status standard refers to the system used by the IUCN to assess and categorize the conservation status of species worldwide. This standard, known as the IUCN Red List Categories and Criteria, classifies species into different categories, such as "Critically Endangered," "Endangered," "Vulnerable," and others, based on factors like population size, range, and threats. The IUCN status standard is widely recognized and used as a global reference for assessing the conservation status of species and guiding conservation efforts.

Standards for the development of APIs:

- REST: REST (Representational State Transfer) is an architectural style for designing networked applications that utilize the HTTP protocol, emphasizing simplicity, scalability, and a uniform interface for resource manipulation.
- OpenAPI: OpenAPI (formerly known as Swagger) is a specification and set of tools that enables the design, documentation, and interaction with RESTful APIs in a standardized and machine-readable format.
- ARISE digital species identification API: the ARISE project has defined a standardised way for species identification algorithms to describe what they have localized and identified in media items. It is flexible enough to cover images, sounds, video, and more; it can represent general classification results as well as detections in "bounding boxes", and detections linked across multiple images.

Standard used for the sharing of maps and geospatial data:

- OGC (WMS, WFS): OGC (Open Geospatial Consortium) is an international organization that develops open standards for geospatial data and services. WMS (Web Map Service) and WFS (Web Feature Service) are two OGC standards: WMS provides a way to serve and request map images over the web, while WFS enables the exchange and querying of geospatial feature data.
- GeoTiff and Cloud Optimized GeoTIFF: GeoTIFF is a widely used standard file format that combines geospatial information and raster imagery, allowing for the storage of geographic coordinates, map projections, and other spatial metadata within the image file. Traditional GeoTIFF files store raster data in a way that can be challenging to work with in cloud environments due to their large file sizes and the need to read entire files for accessing specific portions of the data. Cloud Optimized GeoTIFF, on the other hand, are designed to address these challenges using tiling, HTTP access, and internal overviews to optimise performance. COGs enable streamlined processing and visualisation of geospatial data in cloud-based applications.
- NetCDF and EBV NetCDF: NetCDF (Network Common Data Form) is a file format and software library commonly used in scientific computing. It's designed to store multidimensional data arrays, making it suitable for climate, oceanography, and other scientific data, typically

rasters. NetCDF files enable efficient data storage, access, and sharing across different platforms and programming languages. EBV NetCDF is a particular hierarchical structure of a netCDF file designed to hold Essential Biodiversity Variables. The structure allows several data cubes per netCDF file. These cubes have four dimensions: longitude, latitude, time and entity, whereby the last dimension can, e.g., encompass different species or groups of species, ecosystem types or other. Each cube holds data of a specific metric. The usage of hierarchical groups enables the coexistence of multiple data cubes, each sharing the same dimensions.

- GeoJSON: GeoJSON is a file format for encoding geospatial data structures using JSON, allowing for the representation of points, lines, polygons, and other spatial features in a lightweight and interoperable manner.
- Geopackage: GeoPackage is an open, standards-based, platform-independent, portable, self-describing, compact format for transferring geospatial information. GeoPackage offers the advantage of a single, self-contained file structure that supports both vector and raster data, simplifying data management and promoting cross-platform compatibility, making it a versatile and user-friendly geospatial data format.
- INSPIRE metadata: INSPIRE metadata is a standardized set of metadata specifications and guidelines developed by the European Union's INSPIRE Directive, aiming to facilitate the discovery, evaluation, and access to geospatial data resources across Europe, ensuring interoperability, harmonization, and the sharing of environmental and spatial information.
- LAS/LAZ: The LAS format is a file format designed for the interchange and archiving of LiDAR point cloud data. It is an open, binary format specified by the American Society for Photogrammetry and Remote sensing (ASPRS). The format is widely used and regarded as an industry standard for LiDAR data. The LAZ file format is a compressed LAS file format, which is typically less than 20% of the corresponding LAS file in size and thus can more easily be stored and shared among others. Both of the two formats are supported currently with LAStools.
- Shapefiles: The Shapefile format is an Esri vector data storage format for storing the location, shape, and attributes of geographic features. It is stored as a set of related files and contains one feature class. It is developed and regulated by Esri as a mostly open specification for data interoperability among Esri and other GIS software products. The specification of the format can be found here.

Other generic standards used:

- json (metadata , URLs, textual data, numbers, etc.)
- Comma-separated (CSV), or tab-delimited (TAB), xls (for the export of tabular data)
- FlatBuffer for the sharing of compressed version of big data
- JPG, PNG and exif for images
- MP3, WAV for sound
- MPEG, MP4, MOV, AVI for movies
- ISO-639-1/2/3 for languages (default language: EN)

3 List of MAMBO solutions and related plans in terms of platforms and formats

3.1 Light trap (AMI-trap) - CEH

3.1.1 Solution description: The AMI-trap consists of Ultra-violet and white lights for attracting and imaging moths, high-capacity data storage to collate images over long sampling periods, battery and solar power to allow the system to be deployed away from mains power, and customisable sampling schedules.

Images collected can be processed through your own workflow, or using the AMI-trap Data Companion (under development by our partners at eButterfly), which has existing classifiers for the UK and Denmark, as well as Vermont and Montreal (under development by partners at Mila Quebec AI Institute). This tool will find moths in the images collected and try to identify them to species, giving the species name, as well as the uncertainty of the predication.

3.1.2 Type of data collected/produced: The AMI trap produces JPEG images of the moth screen. Meta data is added to these including location, data and system configuration.

3.1.3 Plans in terms of platforms and formats: Work is ongoing to refine the metadata standards to use for AMI images. This includes a review of existing standards to understand what already exists, and what new fields might be needed to account for this new type of data. The AMI trap itself is continually under development, with future advances to include the addition of acoustics, remote systems health checks, and camera upgrades.

3.2 Light trap (Manual bucket traps) - CEH

3.2.1 Solution description: The moth 'LED bucket trap' is a low cost, Do-It-Yourself, hardware setup for capturing moths. Made from cheap components that are easy to buy at a local hardware shop the bucket trap design offers an easy way for people to start monitoring moths. This system is used for a monitoring programme involving farmers in the Netherlands and is a protocol being used by the SPRING project that is piloting an EU Pollinator Monitoring Scheme following Potts et al. (2020).

3.2.2 Type of data collected/produced:

The 'LED bucket trap' attracts moths overnight that are then counted the following morning, using image analysis (the NIA classifier as used in ObsIdentify and Observation.org) to help identify species. The trap location is recorded in terms of spatial location (lat., long.), alongside date. Moth populations can change over year as new species emerge and build up in numbers - it is therefore recommended that populations are sampled regularly through the main period of interest (e.g. summer months). For MAMBO, we recommend at least monthly sampling between the beginning of April and end of September - more frequent sampling is also welcome.

3.2.3 Plans in terms of platforms and formats:

The power of monitoring is in repeated measurements, and following the same protocol so densities can be compared in space and time. After fieldwork making the counts available is a vital next step. The data can be entered by registering a trap location at https://butterfly-monitoring.net/my-moth-traps and use either this website to enter counts, or the ButterflyCount app (available for Android and iPhone). That way data can be immediately used for research and conservation, and also downloaded as an excel file for personal use.

3.3 Pollinator camera - AU

3.3.1 Solution description: We will use two types of pollinator cameras. Across demonstration sites, we will work with commercially available low-cost time lapse cameras (Wingscapes and TimeLapsePro) mounted on sturdy metal frames facing the ground. These cameras have a

maximum recording frequency of one image every 10 seconds. The cameras can be powered by eight AA batteries internally or externally via a solar panel and 12V car battery. With external power, up to six cameras can be powered from a 60W solar panels during summer at the planned latitudes of deployment. We will combine these cameras with standard flower beds to minimize the challenge of detecting insects against complex backgrounds. A second, more complex camera type will be used to test edge detection and tracking of insects at one site. This camera type uses a USB web camera, stores images on an SSD hard drive and involves processing images using a GPU enabled mini-computer such as JetsonNano.

3.3.2 Type of data collected/produced: The pollinator camera produces jpeg images of the vegetation and insects below the camera. The camera has a built-in temperature sensor and temperature at recording is stored in the exif data of each image. Metadata includes location, camera name, additional information about surrounding plant communities and recording schedule.

3.3.3 Plans in terms of platforms and formats: The pollinator camera with edge detection capabilities is undergoing further development, with future functionality likely to include remote data transmission and camera upgrades.

3.4 Audio ID API - Acoustic detection and monitoring of animals- NATURALIS

3.4.1 Solution description: For European breeding birds, bats, grasshoppers and marine mammals sound recognition models are being developed. These models will be freely available and can be deployed and used by institutes throughout Europe. Deploying these models takes expertise and budget. In order to allow as many end users as possible to access the models we also aim to make them available through several biodiversity portals in Europe. The models can be used either for the analyses of long recordings (for instance for monitoring) or for the analyses of short recordings by citizen scientists. For the latter group of end users we aim to make the model available through the same cooperation of European biodiversity portals as described for image recognition. For users wishing to analyse long recordings (hours, days or even weeks long) we currently aim to provide a service through the Naturalis based infrastructure Arise. This will be tested in winter 2023-2024 and when successful we will

develop a structure allowing end users to access the models for a fee which will cover the cost of deployment and maintenance.

3.4.2 Type of data collected/produced: The input data for the sound recognition consist of audio files (wav, mp3). Connected to these is metadata regarding the time and location of the recordings. The datatypes and structure follow DarwinCore. Records resulting from the use of the sound recognition models will be stored in the connected European Biodiversity portals (see image recognition) and will annually be uploaded to GBIF. Analyses of long recordings will likely be arranged through Arise, where the output will be JSON data following a specified format in compliance with DarwinCore. The further processing of the data is dependent on the different end users.

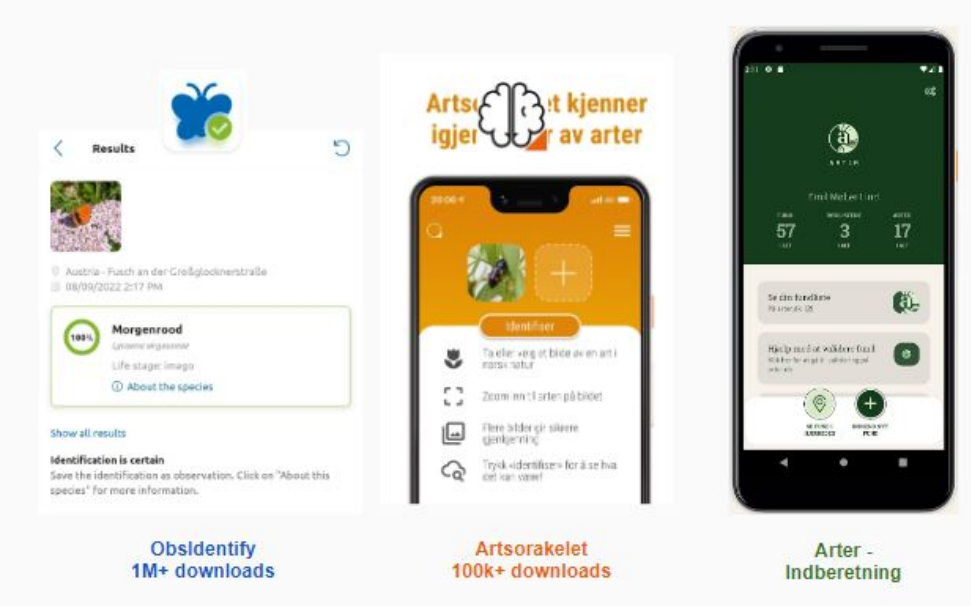
3.4.3 Plans in terms of platforms and formats:

In winter 2023-2024 the recognition model for European breeding birds will be taken into test-production as part of the MSM model cooperation of the European Biodiversity portals. Depending on the results and the feedback of the different cooperating portals it will be decided whether or not this deployment will be maintained in the long run and if it will be expanded to grasshoppers and bats. Our main focus is on developing the sound recognition as a tool for the analysis of long recordings used for biomonitoring. In order to support this the bird recognition model will be deployed within the Arise infrastructure in 2024. Based on several test projects for EU Horizon projects MAMBO, TETTRIs and GUARDEN it will be decided if the Arise infrastructure is indeed the best place to deploy and maintain this service. If so, a service structure needs to be developed which gives end users access to the models for a fee covering the cost of deployment (storage and compute capacity).

3.5 Animal Photo ID - AI based image recognition for European animals - NATURALIS

3.5.1 Solution description: The products of this project consist of AI-based image recognition models that can be taken into production by various partners. Taking models into long term production and maintaining the service takes expertise and budget. For this reason, we are constructing a cooperation between European biodiversity portals aiming at sharing training data and splitting the cost involved for the long term maintenance. This cooperation is making use of the model building architectures and models made in the MAMBO project. In 2023 this

cooperation resulted in the first large scale European model: Multi Source Model (MSM) which supports the identification of over 30.000 European species of fungi, plants and animals. This model is currently being applied by the following end users: Artsdatabanken Norway (NIBC, Norway), SLU Artdatabanken (SSIC, Sweden), UK Centre for Ecology & Hydrology (CEH, UK), Arter (Denmark), Finnish Museum of Natural History Luomus (Laji, Finland), Senckenberg (Germany), Waarneming.nl (Netherlands), Waarnemingen.be – (Belgium). Of these the Observation.org portal is the only one available throughout Europe and it allows citizen scientists from all European countries to access the service provided by the MSM.



Examples of three apps connected with European biodiversity portals which since 2023 allow citizen scientists to access the service provided by the MSM. ObsIdentify is connected to Observation.org and is available throughout Europe while Artsorakelet and ArterIndberetning are connected to the Norwegian and Danish biodiversity portals.

3.5.2 Type of data collected/produced: The input data for the MSM is an image of a fungi, plant or animal taken in the field anywhere in Europe. The image has to be accompanied by data on location and the date. After processing the MSM gives species predictions accompanied with certainty scores. The users then decide if these are valid and, based on this, decide to upload the image and the metadata as a record. All cooperating biodiversity portals regularly upload

their data to GBIF and as such apply the data standards set by GBIF (see above). The image, the metadata and information on which AI-model was used is stored in the datafiles of the different biodiversity portals.

3.5.3 Plans in terms of platforms and formats: The biodiversity portals cooperating on the MSM collect tens of millions of records annually. It seems likely that an increasing amount of these records will at least be partly connected with the AI image recognition provided by the MSM either through the original identification or in the validation process (the model supporting the quality check of the identifications). While all portals utilising the MSM comply with GBIF's data standards and upload their data accordingly, there remains a necessity for further standardisation. Specifically, there's a need to establish a more comprehensive tracking system for AI image recognition across diverse biodiversity portals. This would enable the precise tracing of utilised model versions for each record and the decisions derived from image recognition.

3.6 Pl@ntNet app - AI-assisted collection of plant observations and plot surveys - CIRAD / INRIA

3.6.1 *Solution description:* Pl@ntNet is a citizen science platform and information system that relies on Artificial Intelligence (AI) to facilitate the identification and inventory of plant species. It has three main front-ends: Pl@ntNet androïd, Pl@ntNet iOS and Pl@ntNet web. Pl@ntNet is based on a cooperative learning principle: (i) the Pl@ntNet user community generates a large number of plant observations in the field. An observation can contain from one to five images of the observed specimen and a set of metadata such as the date of observation, the GPS (Global Positioning System) coordinate, or the type of view of image (flower, leaf, fruit, entire plant, etc.), (ii) each observation is automatically identified by an AI algorithm (detailed later in this section) and the candidate recognized species are returned to the observer for confirmation. All observations are stored in a database and can be reviewed by the community. A user can typically confirm the initial species name shared by the observer (with the help of the AI) or propose an alternative species name. Only the observations that reach a sufficient degree of confidence are labelled as valid observations and are added to the training set of the AI model. To take into account the new observations and revisions made by the community, the model is retrained regularly. This is typically done on a monthly basis

(rather than, e.g., a daily basis) to limit energy consumption and carbon emissions. Within MAMBO (and in collaboration with the EU project GUARDEN), Pl@ntNet will be extended so as to process vegetation plots records in addition to the individual plant records as is currently the case. A user will be able to create an observation plot with associated metadata (title, description, date), to populate it with one or more HD image of the plot and then to automatically analyse the species present in the HD image(s) based on the AI model developed in task T3.4 of MAMBO. The model will allow estimating:

- the set of species present in the plot
- the coverage percentage of each species in the HD image(s)
- the position where each species is detected in the HD image(s)

3.6.2 *Type of data collected/produced:* each vegetation plot will include metadata (geo-location, date, author, licence, species list, species coverage, etc.), HD images of the plot, and, optionally, links to individual Pl@ntNet observations of the specimens present in the plot.

3.6.3 *Plans in terms of platforms and formats:* classical Pl@ntNet observations as well as vegetation plot data will mainly be managed within the Pl@ntNet platform itself. Each individual observation and plot survey will be assigned a unique resource identifier (hosted on https://identify.plantnet.org/) and a creative common licence (cc-by-sa). The plot surveys will be shared via new dedicated views in the Pl@ntNet application (on the web version initially). As for classical Pl@ntNet observations, plot survey data of sufficient quality will eventually also be published in GBIF (through a new dataset managed by Pl@ntNet as GBIF data provider).

3.7 Pl@ntNet API - an open API for the integration of Pl@ntNet AI-based services in external applications - CIRAD-INRIA

3.7.1 *Solution description:* Pl@ntNet API provides a computational access to the visual identification engine used in Pl@ntNet apps in the form of a RESTful Web service. The service allows users to submit from 1 to 5 images of an individual plant and to have in return the list of the most likely species as well as a confidence score for each of them. The identification engine is based on most advanced deep learning technologies (Vision Transformers) and is regularly updated thanks to the contributions of the community and the integration of new expert databases. Within MAMBO (and in collaboration with the EU project GUARDEN), Pl@ntNet API will be

extended so as to process vegetation plots records in addition to the individual plant records as is currently the case (based on the results of task T3.4). Moreover, a new service dedicated to the location-based prediction of species and habitats will be developed based on the results of task T4.1 of MAMBO (at EU scale and high spatial resolution).

3.7.2 *Type of data collected/produced:* By default, Pl@ntNet API does not store the data sent to the API. It is up to the developers of the external applications to manage the data they sent and the data produced with the API according to their own data management plan. For particular institutes having a "partner" role within Pl@ntNet, it is possible to contribute data to the database. In such cases, the sent data is managed in the same way as classical Pl@ntNet observations collected via the Pl@ntNet app (see section 3.6).

3.7.3 *Plans in terms of platforms and formats*: Pl@ntNet API is hosted on Pl@ntNet infrastructure (at CIRAD) and is publicly accessible at https://my.plantnet.org/. It relies on a json based format aligned with Pl@ntNet data model. New routes will be developed to be compliant with the ARISE digital species identification API format to foster interoperability with other EU projects related to biodiversity monitoring. Pl@ntNet API is already integrated in EOSC marketplace (see here) and will be updated with the new services and routes developed.

3.8 Deep-SDMs - very high-resolution mapping of species and habitats based on remote sensing, citizen science data and deep Species Distribution Models - CIRAD-INRIA

3.8.1 *Solution description:* in the context of task 4.1 of MAMBO, we develop new deep learning models (called Deep-SDMs) for the mapping of species and habitats at very high resolution. Those models couple large volume of available plant species occurrences and occupancy data (e.g, GBIF, sPlot, Pl@ntNet, Euroveg) with RS landscape patterns (using satellite ortho-imagery) and other environmental spatial data to train deep neural networks with high predictive power. The models are trained through an open source framework called malpolon.

3.8.2 *Type of data collected/produced:* The training of Deep SDMs first requires aggregating, cleaning and formatting large training datasets composed of heterogeneous types of data (csv files, rasters, images, etc.). The output of the training phase itself is mainly the weights of the

trained models. Finally, the trained models are used to produce maps by applying them in inference mode on a spatial grid (using the same predictors as the training phase).

3.8.3 Plans in terms of platforms and formats:

- The datasets used to train the models will be shared through a public server hosted by Pl@ntNet and through generalist digital repositories such as ZENODO or Dryad (depending on the data volume) when associated with a scientific publication.
- The open source framework used to train the models (malpolon) will be shared on github.
- The trained models will be shared on github or on a public server hosted by Pl@ntNet (depending on the volume).
- The produced maps, in particular the maps of individual plant species (Terabytes), will be centrally stored in Pl@ntNet map server and made available through different tools of the Pl@ntNet platform (depending on scope and resolution):
    - Interactive maps integrated in Pl@ntNet species sheets
    - GeoPl@ntNet, a web application developed in the context of the EU GUARDEN project whose goal is to make biodiversity indicators easily accessible and understandable for everyone in the form of interactive maps and fact-sheets available at EU scale.
    - New dedicated routes in Pl@ntNet-API (location-based prediction in json format, direct maps download based on OGC standards).
    - A generalist map viewer hosted by Pl@ntNet.
    - A selection of maps (e.g. habitat types or other essential biodiversity variables) will be shared through the EBV data portal.
    - Some maps may be shared through ZENODO and DRYAD when associated with scientific publications requiring a DOI.

3.9 Airborne LiDAR - workflows and data products for habitat condition metrics from light detection and ranging

3.9.1 Solution description: In the context of task T4.3 and task T6.2, LiDAR metrics for habitat condition assessments will be calculated from massive point cloud datasets that are openly accessible from repositories of national airborne laser scanning flight campaigns. Processing of the LiDAR raw data will be done using Free Open Source Software (FOSS) such as the cross-platform Python tool Laserchicken and the high-throughput workflow from Laserfarm. The

resulting data products of ecosystem structure will cover LiDAR metrics of vegetation structure, including ecosystem height, ecosystem cover, and ecosystem structural complexity variables.

3.9.2 *Type of data collected/produced:* The raw data (point clouds) from national airborne laser scanning flight campaigns are downloaded to a local data storage. After processing with the Laserfarm workflow, the resulting data products will contain LiDAR metrics capturing ecosystem structure in three key dimensions (ecosystem height, ecosystem cover and ecosystem structural complexity). An example of such a data product of ecosystem structure variables at 10 m resolution across the whole Netherlands is available here.

3.9.3 *Plans in terms of platforms and formats:* The point clouds from national airborne laser scanning flight campaigns are typically available in LAS/LAZ format. After processing, the data products of ecosystem structure will be raster layers in GeoTiff format. These data products will be made openly accessible via the Zenodo digital repository. Data processing will mostly be done using the computing services (e.g. virtual machines) and storage systems from the Dutch national IT infrastructure SURF.

3.10 Drones monitoring

3.10.1 *Solution description:* In the context of task T4.2, LiDAR metrics describing habitat conditions will be calculated using UAV photogrammetry and/or UAV LiDAR. The data will be collected using the DJI Zenmuse L1 and DJI Zenmuse P1 system for LiDAR data and photographs, respectively. For the purposes of georeferencing of the products, the GNSS data will be collected in the form of RTK and static measurements. The processing of raw data will be performed using OpenDroneMap open source software (for generation of point cloud from dense image matching) and DJI Terra (for processing of LiDAR data). Habitat condition metrics will be calculated based on UAV products (georeferenced point clouds, orthomosaics). For this purpose, deep learning methods will be developed and implemented in Python. The results will be validated using in situ measurements.

3.10.2 *Type of data collected/produced:*

For LiDAR the raw data collected are point clouds, for RGB imagery raw data are individual image frames that are processed to produce 2 dimensional RGB image mosaics, point clouds and digital surface models. The data produced will be 2- dimensional rasters of habitat condition metrics, text files with classification results and point clouds.

3.10.3 *Plans in terms of platforms and formats:* The products of UAV measurements will include: 1) georeferenced LiDAR point clouds stored and distributed as LAS/LAZ files, 2) georeferenced photogrammetry point clouds stored and distributed as LAS/LAZ files, 3) georeferenced orthomosaics stored and distributed in GeoTIFF format. After deep learning based processing, the results will be stored and distributed as text files, LAS/LAZ files and raster layers stored as GeoTIFFs. The data products will be made openly accessible via the Zenodo digital repository.

4 Conclusion

This deliverable provides guidelines about the data standards and technical specifications to be used by MAMBO tools so as to make them FAIR (Findable Accessible Interoperable and Reusable) at the end of the project. The document highlights the recommended open data platforms for managing and sharing project-produced biodiversity data and hosting the developed tools. It subsequently outlines the recommended standards for data storage and exchange formats, followed by providing detailed guidance on the innovative technologies developed as part of the MAMBO project. It is important to note that this roadmap is not a specification for a complete integrated system. MAMBO is based on an agile approach in which each technology is developed independently, but relying on existing open data platforms and standards that facilitate cost-efficiency, scalability, interoperability, community support, security and reliability.

MAMBO

MODERN APPROACHES TO THE MONITORING OF BIODIVERSITY

www.mambo-project.eu

Project partners

AARHUS UNIVERSITY

Naturalis Biodiversity Center

UFZ

University of Reading

Inria

PENSOFT

UK CEH

cirad

ECOSTACK INNOVATIONS LIMITED