Project Title	Biodiversity Digital Twin for Advanced Modelling, Simulation and Prediction Capabilities
Project Acronym	BioDT
Project Number	101057437
Type of Project	RIA - Research and Innovation Action
Topics	HORIZON-INFRA-2021-TECH-01-01
Starting Date of Project	1 June 2022
Ending Date of Project	31 May 2025
Duration of the Project	36 months
Website	www.biodt.eu

D4.1 - Report on the Availability of Consolidated, Improved, Fit for Use Data Streams, Data Model Availability, Improvement, and Application

Work Package	WP4 | Data and Use Cases
Task	T4.2 | Data streams
Lead Authors	Dmitry Schigel (GBIFS)
Contributors	Taimur Khan (UFZ), Kyle Eyvindson (NMBU), Ossi Nokelainen (JYU), Jan Dick (UKCEH), Chris Andrews (UKCEH), Desalegn Chala (UiO), Erik Kusch (UiO), Carrie J. Andrew (UiO), Dag Endresen (UiO), Kate Ingenloff (GBIFS), Tobias Frøslev (GBIFS), Hanna Koivula (CSC)
Peer Reviewers	Franziska Taubert (UFZ), Sharif Islam (Naturalis),
Version	V1.0
Due Date	31/05/2024
Submission Date	31/05/2024

Dissemination Level

X	PU: Public
	SEN: Sensitive – limited under the conditions of the Grant Agreement
	EU-RES. Classified Information: RESTREINT UE (Commission Decision 2005/444/EC)
	EU-CON. Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC)
	EU-SEC. Classified Information: SECRET UE (Commission Decision 2005/444/EC)

Funded by the European Union

Version History

Revision	Date	Editors	Comments
0.1	06/03/2024	Dmitry Schigel (GBIFS)	Draft of Table of Content
0.2	02/05/2024	Dmitry Schigel (GBIFS)	Version 0.2
0.3	16/05/2024	Dmitry Schigel (GBIFS)	Version 0.3, resolving comments from two project revewiers
1.0	30/05/2024	Gabriela de Paula Souza Zuquim (CSC), Jenni Poutanen (CSC)	Submission-ready version incorporating BioDT Project Management Office (PMO) editorial changes

Glossary of Terms

Item	Description
DWD	Deutscher Wetterdienst (German Weather Service)
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	The fifth generation ECMWF atmospheric reanalysis of the global climate covering the period from January 1940 to present. ERA5 is produced by the Copernicus Climate Change Service (C3S) at ECMWF.
HMSC	Hierarchical Modelling of Species Communities
NFI	National Forest Inventory
oToL	Open tree of life https://tree.opentreeoflife.org
pDT	Prototype digital twin

Keywords

Biodiversity, Digital Twin, Modelling, Simulation, Prediction, High-Performance Computing

Disclaimer

The BioDT project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101057437 (https://doi.org/10.3030/101057437). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Executive Summary

A variety of streams of biodiversity data from a range of sources and scales, as well as of accompanying abiotic data is gathered and developed, through interactions with RIs and expert communities, to ensure the combining the BioDT functions with FAIR attribution and data citations. This report covers six data aspects of the BioDT project across ten prototype digital twins (pDTs). These aspects are 1) data collection, 2) data location, 3) dataset and model description, 4) data structure, 5) data updates, and 6) data sharing. In addition, biodiversity and non-biodiversity data are covered in the respective sections.

The ten prototype digital twins (pDTs) cover four themes of different data complexity ranging from fundamental biodiversity explorations and predictions to modelling biodiversity changes in the context of human life. The groups are: **Thematic group 1** - Species response to environmental change: Grassland biodiversity dynamics, Forest bird biodiversity dynamics, Real-time bird monitoring with citizen science data, Cultural ecosystem services; **Thematic group 2** - Genetically detected biodiversity: Crop wild relatives and genetic resources for food security, DNA detected biodiversity in cryptic habitats – prioritisation of sampling effort & phylogenetic diversity; **Thematic group 3** - Dynamics and threats from and for species of policy concern: Invasive species; **Thematic group 4** - Species interactions with each other and with humans: Disease outbreaks, Pollinators. The data streams aspects are detailed in the main text of the report.

Expectedly, handling data across the pDT have demonstrated differences in all the six aspects that we have explored. These differences are reflected not only in the unevenness in data availability around individual topics across use cases (such as species distribution, species status, environmental, and land use data) but also in the different backgrounds and data experiences in the teams prototyping the digital twins. Overall maturity of the data streams is higher in the teams with strong prior data exposure the FAIRness than in the groups where expertise is dominated by modelling, algorithms and ensuring technical solutions. While this may be natural, it is an important lesson learnt that may guide following projects – to pay attention to data culture and data expertise in addition to data availability and data access technical solutions.

Project´s high ambition to prototype as many as ten prototype use cases divided in four groups has certainly contributed to the observed data streams heterogeneity, which is an operational challenge, but also a broad learning opportunity in the project.

Overall, all ten pDTs have shown progress and shift from initially only locally available test data to or towards versioning at or in connection to supercomputer, satisfying project internal data needs. As project is now entering the third and final project year, and, at the same time, pDTs generally move from the construction to improvement and finalization stage, it is an optimal moment to improve FAIRness and global open access to the data resources, ensure dynamic and automated updates of data sources is in place, and that availability of data resources improves beyond the project scope and timelines in the EU and worldwide.

Table of Contents

1. Availability of data streams

Each of the individual prototype digital twin teams have been given freedom to identify and apply the necessary data that are fit for the prototyping purposes. These data availability efforts under Task 4.2 forked into accessing biodiversity data (section 1.1, at the core of this project) and non-biodiversity data (section 1.2), and these initially ranged from hypothetical data to locally available test data to availability at the institutional level to global research infrastructures (RIs), with on-demand or constant (API based) access, and, eventually, in some cases, to mirroring these data streams at the EuroHPC LUMI supercomputer (https://www.lumi-supercomputer.eu).

The mid-project status of data availability, the source for this deliverable report, has been collected through a series of data interviews in collaboration with WP5 (Sharif Islam, Naturalis). The following standard interview protocol was used, each of ten interviews took place over Zoom for 25 minutes with the data streams contacts identified in each of the pDT teams. The respondents were given an additional time of 1 week to complete, correct, and update their responses.

The following interview protocol has been used:

1.1 Data Interview Protocol

Background

We are committed to ensuring that BioDT adheres to the FAIR principles, encompassing the entire lifecycle from generating, re-using data and models to the final outputs. The complexity arises from the diverse nature of data sources, models, software, and tools used in developing the pDTs. There is variation and divergence within and across pDTs, demanding a delicate balance between productive progress and alignment towards FAIR and the overall Digital Twin vision.

In the formal context of the project implementation, there are at least two anchor points to pay attention to: Task 4.2 (Data streams) and WP5 (Improving Quality of Data, Workflows and Models through FAIR Data Principles)

For convenience, relevant key lines to describe the Task 4.2 and WP5 are inserted below:

T4.2 This task will gather, translate and combine the existing European and global biodiversity data necessary for carrying out the modelling applications across the above use cases. It will re-use and combine data products and services from various sources including GBIF.org (EOSC service). Datasets needed for the models and applications will require a new combination of existing biodiversity and abiotic data, increased data quality and information confidence and processes (in collaboration with WP5) to ensure the right fit for purpose in terms of data formats or data mobilisation.

The aim of WP5 is to facilitate the interoperability of the Biodiversity Digital Twin, particularly the modelling processes by adding quality indicators, improving the FAIRness and the reusability of the datasets, services and models. In this work package, we will: consolidate the FAIR data designs of the involved Research Infrastructures (RIs) through a layer of FAIR Digital Objects (FDO); pilot semantic mapping to improve interoperability and reusability of semantic artifacts; apply FAIR Digital Objects to the services; and research outputs for improved reusability and develop a framework to indicate data quality as part of the data product.

Objectives

- Review the status of the data streams / data dependencies for each pDT
- Re-emphasise the importance of FAIR principles within the context of BioDT
- Identify the current state of each pDT and ascertain the alignment with FAIR principles

- Determine specific areas where assistance is required to bridge the gap between current practices and the FAIR vision.

Questions to guide the interview process

These questions served two purposes. Firstly, they help in understanding the intricacies of each pDT's data landscape. Secondly, the insights garnered from these responses will contribute to the development of common approaches (standardisation, creating high-level metadata profiles: https://github.com/BioDT/biodt-fair/discussions/3).

1. **Data Collection** (if this is relevant; when new data is being generated/collected for example from a sensor)
    a. Where and when was the data collected?
    b. Was there a specific domain standard or procedure followed during data collection?
2. **Where is the data?**
    a. Where and how these collected data stored? Add links as much as possible., identify access protocols in case data is not OA.
3. **Dataset and Model Description** (metadata)
    a. How many datasets does the model utilise? Are these datasets described? If not: please create if possible. Such descriptions will help create a machine-readable metadata profile.
    b. Who can add metadata descriptions, if missing?
    c. What is the size of these datasets, and do they vary significantly?
4. **Data Structure**
    a. Describe a typical data file (input file for instance; is it an image, a text file, or a combination of different files). Ideally, provide an example of input data.
    b. Is the data format standardised (such as xml, json, csv) and structured?
    c. No data is self-explanatory. Does metadata serve sufficiently as readme?
5. **Data updates**
    a. Are the datasets expected to be updated with more recent data, and how frequently does this occur?
    b. Is versioning well-documented and transparent, how is it ensured?
6. **Data sharing**
    a. Are there any conflicts or challenges in making some of the data open publicly?
    b. Are there any licensing issues?
    c. What, if any, persistent identifiers are used at the dataset and possibly other levels?
    d. Are there APIs available?
7. **Assistance**
    a. Are there specific challenges or roadblocks hindering the integration of FAIR principles into your current workflow?
    b. If help is welcome, when and how do you prefer to receive it?

1.2 Data Collection

pDT teams interpreted "data collection" in a number of ways – from documenting field methods used in data generation to approaches of data retrieval from the online sources, with the dominance of the latter.

1.2.1 Grassland Biodiversity Dynamics

In this pDT, data spatial scale covered selected locations in Europe for grassland simulations with different temporal scales: daily for weather data, representative snapshots of certain time points for soil and land management. The data sources were retrieved from Copernicus, SoilGrids250m, Pangea (land use maps), specific websites (other maps), see details in Appendix I. Frequency of retrieval is determined by the model prototyping needs, and potentially reacting to updates when public sources are updated. Python scripts have been developed for data retrieval via API. See details described in BioDT Deliverable 6.1 on pipeline scripts for obtaining data (section 2.1.1.1.4). Deliverable 6.1 will be made available in Autumn 2024 after EC approval and will be found in BioDT webpage (https://biodt.eu/documents-publications) and in Cordis (https://cordis.europa.eu/project/id/101057437/results).

1.2.2 Forest/bird Biodiversity Dynamics

This pDT used specific data sources for individual simulations: for LANDIS-II forest simulations, climate data was sourced from Earth System Grid Federation, including averaged, monthly data from the last 30 years and projections of future monthly changes according to two different radiative forcing scenarios (RCP 4.5 and RCP 8.5, three different climate scenarios in total). For National Forest Inventory data, the Natural Resources Institute in Finland (LUKE) was sourced, and for eco-regions 11 active eco-regions are determined based on combined CORINE Land Cover maps and soil maps (Land cover data: CORINE Land Cover, soil data: European Soil Data Centre (ESDAC).

Separately, for the HMSC model species occurrence data were retrieved from the Finnish Museum of Natural History (LUOMUS, which is also FinBIF – a Finnish national node of GBIF – BioDT project's partner RI). The species data covers the years 2007 to 2019 and includes a total of 2920 transects for which the occurrences of 190 bird species were recorded. Species trait data are based on life-history characteristics of European birds (https://doi.org/10.1111/geb.12709).

1.2.3 Real-Time Bird Monitoring with Citizen Science Data

This pDT is based on a number of data sources. Some of these are acquired from existing data sources while others are acquired actively over the project duration. These include land cover (CORINE), forest structure (NFI), climate (COPERNICUS), bird data (in three variants: line transects, species occurrence data from Laji.fi, and LIFEPLAN project recordings), and weather data (ECMWF).

1.2.4 Cultural Ecosystem Services

This pDT has been initially parameterized by existing datasets. Through adaptive sampling we will identify areas where we need biodiversity data but direct them towards iRecord/iNaturalist/NESBREC data portals which will end up in GBIF and feed into the digital twin after some time lag. The data are opportunistically recorded. User will also personally score aspects of the recreation elements which will result in updated DT output specific for them. Each input dataset has its own documented standards.

1.2.5 Crop Wild Relatives and Genetic Resources for Food Security

This pDT depends on occurrence data from https://www.gbif.org/occurrence-snapshots (as for Jan 2024: >2.6B). Climate data: ERA5 - ECMWF 's reanalysis combining model data and observations for global climate and weather from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels. Soil parameters such as soil moisture and soil temperature at different depths: ERA5-Land is a reanalysis dataset from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land-monthly-means. All data were retrieved on global scale. The GBIF snapshot combines occurrence data from a variety of field protocols

for data collection harmonized to TDWG standards (DwC, ABCD). ERA5 datasets where resampled to 9 km resolution.

1.2.6 Prioritizing Future eDNA Sampling

This pDT is also based on the global taxon occurrences (2.6+ billion) from GBIF.org, potentially including all species occurrences GBIF in the 90.000+ datasets mediated and standardized through GBIF.org (no delimited time or place of sampling). Initially the pDT is limited to eDNA-based biodiversity data however, which is a subset of the GBIF data. eDNA data which by definition follows protocols suitable for that kind of material (DNA extraction, PCR, sequencing, bioinformatic processing), but given that, GBIF does not enforce any standardised procedures, but accept any data derived using those overall steps.

1.2.7 Phylogenetic Diversity

This pDT also rely on global taxon occurrences (2.6+ billion) from GBIF.org. Potentially including all species occurrences GBIF in the 90.000+ datasets mediated and standardized through GBIF.org, meaning that there is no delimited time or place of sampling. Global phylogeny data are harvested from oToL (Open tree of life: https://tree.opentreeoflife.org) without time or space limitations. It is based on aggregating data from available sources. Species occurrence data in GBIF is standardized according to the Darwin Core Standard (https://dwc.tdwg.org), but this is not related to original data collection procedure. Open Tree of Life constructs a comprehensive, dynamic, and digitally available tree of life by synthesizing published phylogenetic trees along with taxonomic data. The published (contributed) smaller trees may be constructed from data collected in any way suitable to construct a phylogenetic estimate.

1.2.8 Invasive Alien Plant Species

For the Invasive Alien Species pDT, a set of initial variables was processed into the reference grid (10 km × 10 km) in a ready form to be used in the models or for post-processing of model outputs using FAIR principle. Climatological data sourced: CHELSA (time span 1981–2020; spatial resolution 30 arc seconds). CHELSA provides current and future (5 climate CMIP6 models × 3 shared socioeconomic pathways [ssp126 - ssp370 - ssp585] × 3 time periods [2011-2040; 2041-2070; 2071-2100]) projections for 46 variables. All these variables were rescaled, and a small subset of less-correlated predictors will be used in the models.

Habitat type. Source: CORINE Land Cover (Copernicus; spatial resolution 100 m). CORINE data included broad habitat types and percentage coverage calculated for each type per 10 km × 10 km. Road and railway density was taken from Global Roads Inventory Dataset and OpenStreetMap. Measured as the total length (in km) of roads/railways per grid cell and represents site accessibility (i.e., describe spatial sampling bias in the GBIF mediated data), the level of habitat disturbance, and the likelihood of dispersal for invasive plants. Sampling effort was sourced GBIF and measured as the total number of GBIF occurrences per grid cell for all vascular plant species observed after 1980. It will be used to correct spatial sampling bias in the opportunistic presence only GBIF data.

For response variables, the most recent checklist of naturalized alien plant species of non-European origin for Europe were obtained from FloraVeg.EU. We standardized this list against the GBIF taxonomic backbone and selected species with a sufficient number of occurrences. The following data sources were used to assemble the occurrence dataset: 1) GBIF, for which we used the *rgbif* R package to download occurrence data for each species (a total of >10 million occurrences, March 2023). Doubtful occurrences or occurrences with high spatial uncertainty were excluded; 2) The European Alien Species Information Network (EASIN) which provides data on 14,263 alien species; 3) EASIN data, which is available as presence-only grids at the same resolution and projection that we will use in the models. We standardized EASIN against the GBIF taxonomic backbone and extracted data on the species of interest using EASIN's API. We are particularly interested in non-GBIF data from EASIN (387,547 unique combinations between grid ID and species; March 2023). And 4) eLTER which is a network of sites collecting ecological data for long-term research within the EU. Vegetation data was retrieved >120 sites. Since the sites are from different countries and institutions,

the data is highly heterogeneous and in different data formats. Our workflow homogenizes the data and converts it into the HDF5 format to maintain the data in a binary state for efficient throughput rates. There was no domain standard followed as all data sources are accessible in different formats, however, a standard procedure was developed in the workflows to pull new data and compare with older versions.

1.2.9 Disease Outbreaks

This pDT depends on the wild boar and infection data obtained from https://enetwild.com/data-collection. Landscape data is either user-provided or obtained from eNetWild. The data for our prototype are provided by collaborators at UFZ. There was no standardized data method collection overall, but data is to eNetWild in a standardized format.

1.2.10 Pollinators

The pollinators pDT used input land use data as described in Preidl et al (2016) (https://doi.org/10.1594/PANGAEA.910837). Weather data are accessed from Deutscher Wetterdienst (German Meteorological Service, DWD) via the api *rdwd* (https://cran.r-project.org/web/packages/rdwd/index.html). Calibration data came from: https://www.bienenkunde.rlp.de/Internet/global/inetcntr.nsf/dlr_web_full.xsp?src=7DE6581RTC&p1=510TV6HBBL&p3=5PW3P32TF7&p4=HY3576SY58.

1.3 Data Location

1.3.1 Grassland Biodiversity Dynamics

Original source data is published online (see 1.2.1 and Appendix I), prepared input data (from source data) is stored locally and on LUMI.

1.3.2 Forest/bird Biodiversity Dynamics

For LANDIS-II Forest simulations: climate data: https://aims2.llnl.gov/, National Forest Inventory data: http://kartta.luke.fi/opendata/valinta-en.html. Eco-regions: land cover data: CORINE Land Cover — Copernicus Land Monitoring Service, soil data: ESDAC - European Commission (europa.eu), for the HMSC model: species occurrence data: https://www.luomus.fi/en, species trait data: Dryad | Data Life-history characteristics of European birds (datadryad.org). The processed datasets (input/output) are currently stored in a local currently. They will be stored in LUMI in future.

1.3.3 Real-Time Bird Monitoring with Citizen Science Data

Land cover: CORINE (https://land.copernicus.eu/pan-european/corine-land-cover), Forest structures in CSC Puhti supercomputer (https://docs.csc.fi/computing/systems-puhti/), climate: COPERNICUS (CLIM.1_download_climate_data.R), bird data (type I; Line transects: Obtained from Aleksi Lehikoinen/LUOMUS), Bird data (type II; Laji.fi: S1_download_lajiFI_data.R), weather data (ECMWF: not incorporated), Bird data (type III; LIFEPLAN recordings; note that the stored raw audio data is not intended publicly available. In CSC Allas object storage (https://docs.csc.fi/data/Allas/), downloaded manually by one BioDT member).

1.3.4 Cultural Ecosystem Services

Data location is reported in the UKCEH data management plan.

1.3.5 Crop Wild Relatives and Genetic Resources for Food Security

For the offline modeling: datasets are currently stored on local facilities at the UiO IT-department, setting up pipelines for storing at LUMI cluster is in progress. For the online modeling: Live ingest of data in running simulations will be evaluated.

1.3.6 Prioritizing Future eDNA Sampling

Global taxon occurrences (2+ billion) from GBIF.org. Potentially including all species occurrences GBIF in the 90.000+ datasets mediated and standardized through GBIF.org, meaning that there is no delimited time or place of sampling. Initially the pDT is limited to eDNA-based biodiversity data, which is a subset of the GBIF data. eDNA data which follows protocols suitable for that kind of material (DNA extraction, PCR, sequencing, bioinformatic processing), but given that, GBIF does not enforce any standardised procedures, but accept any data derived using those overall steps.

1.3.7 Phylogenetic Diversity

Data in GBIF is stored in a distributed manner across various participating institutions and networks. These institutions, which include museums, research institutions, and other data holders, provide their data to GBIF. They maintain their own databases and share their data through the GBIF network. This means that the data remains in the custody and management of the institutions that collect and own it, while GBIF acts as a central portal through which this data can be accessed in a standardized form.

The eDNA datasets that the pDT will be built around in the first iteration will be local versions of datasets also available on GBIF.org. Some development is needed to be able to extract the exact data (eDNA based AND specific taxonomic group) directly from GBIF, and potentially reshape it into a format that the model relies on (e.g. OTU table data).

1.3.8 Invasive Alien Plant Species

The processed datasets (input/output) will be stored on LUMI's Object Storage. Invasive DT will be using LUMI's Object Storage (LUMI-O) for model input/output data at each workflow run in this pDT because it offers permanent storage. A clone of select data will be available via a data server built using an OPeNDAP Catalog. LUMI-O gives a public web interface for each individual file in their buckets. Some of the sample data files for certain file formats are uploaded below: CSV: https://465000357.lumidata.eu/iasdt-pub/bzf-catalogue-survey.csv, HDF5: https://465000357.lumidata.eu/iasdt-pub/elter-vegetation.hdf5, RData: https://465000357.lumidata.eu/iasdt-pub/Grid_10_Raster.RData, NetCDF4: https://465000357.lumidata.eu/iasdt-pub/coads_climatology.nc.

One of location issue is that the entire files need to be downloaded to work within third-party systems. The OPenDAP server will clone some defined data from LUMI-O (and MinIO at UFZ for internal usage) into a VM using Docker and will serve it using the Data Access Protocol (DAP), which is a defined data model for accessing remote scientific datasets. The magic here is that DAP allows users to query subsets of the data files, while automatically giving variable-level access, and automatically assigning metadata to the contents of each file. Under-construction documentation: https://khant.pages.ufz.de/opendap/chapters/concept/opendap.html. Template under development: https://git.ufz.de/khant/opendap. The formats under invasives pDT are CSV, HDF5, RData, JSON, and NetCDF file formats for data storage and availability.

1.3.9 Disease Outbreaks

Data is stored locally in a closed repository hosted by our UFZ collaborators.

1.3.10 Pollinators

Data is stored on Pangea, DWD, Bieneninstitute and the BioDT GitHub. Input data: Land use data is described in Preidl et al (2020) (https://doi.org/10.1594/PANGAEA.910837). Weather data are accessed from Deutscher Wetterdienst via the api rdwd: https://cran.r-project.org/web/packages/rdwd/index.html. Calibration data can be found here: https://www.bienenkunde.rlp.de/Internet/global/inetcntr.nsf/dlr_web_full.xsp?src=7DE6581RTC&p1=510TV6HBBL&p3=5PW3P32TF7&p4=HY3576SY58. Storage for the model input data derived from input data and results data under development.

1.4 Dataset and Model Description (metadata)

1.4.1 Grassland Biodiversity Dynamics

Dataset sizes: If datasets are accessed via API, then only small fractions are needed to generate the input data, e.g. input files of size <1 MB, if datasets downloaded. If data is accessed locally, then full maps of size 0-10 GB, fractions are needed and lead to input files of size <1 MB.

1.4.2 Forest/bird Biodiversity Dynamics

Data sizes: when applying pDT to the Uusimaa region of Finland, the size of the processed and prepared input data for LANDIS-II is ~15MB. The size of the processed and prepared input data for HMSC is ~15MB for the same area.

Six different datasets are utilized by two models, LANDIS-II and HMSC. While project runs to add metadata descriptions, if missing, identified BioDT members are in charge of updates with the help of responsible ecologists and modellers for LANDIS-II, and for HMSC.

Real-Time Bird Monitoring with Citizen Science Data

The data will be stored in three datasets (in CSC storage system): **The raw audio data**: ca. 6 million .wav files, each containing sound from 5 secs to 60 mins. The stored raw audiodata is not intended to be publicly available. The audiodata and classification data are retrievable via metadata. **The metadata:** a .csv file where the rows of the table correspond to the audio data files (hence, ca. 6 million rows), and the columns consist of the following information: "user" (anonymized code, unique for each person making recordings), "date" (year-month-day), "time" (hour:minute:second:fractions of second), "len" (as bytes), "real_obs" (real/test recording), "lat" (latitude), "lon (longitude)", "url (location of the file in CSC storage system)", "id" (unique id of the recording), "country" (country where recording was made). **The classification data:** a .csv file where each row is a bird classification (hence, ca. 12 million rows) made by machine learning methods. The columns are "species" (scientific name of the bird species), "prediction" (confidence of the prediction), "rec_id" (the unique id of the recording), "result_id" (the unique id of the classification), "feedback" (True, False or empty depending if user manually verified/falsified the classification or did not do so), "isseen" (whether the user indicated that saw the bird), "isheard" (whether the user indicated that listened to the bird). For updates, identified BioDT members are in charge while the project runs.

Regarding data sizes, the project will generate and handle audio data. E.g., during 2023 we acquired 2,982,008 audio files from which we classified 5,843,492 bird vocalizations. We estimate that the total number of audiophiles will be 6 million and the total number of bird vocalization will be 12 million in 2024.

1.4.3 Cultural Ecosystem Services

Recreation datasets count is 14, for biodiversity 101 (accessed via GBIF), for environmental input 7 datasets. For biodiversity models we will produce a human readable ODMAP report, and machine-readable equivalent. This work is being done by Dylan Carbone and Julian Lopez Gordillo. Further details, the steps in the development and application of the species distribution model are described in The Overview, Data, Model, Assessment and Prediction (ODMAP) in chronological section (Zurell et al 2020: https://doi.org/10.1111/ecog.04960). The protocol is supported by an R shiny application, and users must answer questions related to the model objectives, spatial-temporal extent, data processing, model choice and evaluation, and uncertainty of predictions. The application compiles a human readable report to be included as supplementary information in a publication. Dylan's work integrates the report's compilation as a final stage in the species distribution model workflow, generating many of the answers to the ODMAP questions programmatically from objects in the workflow. The sizes of these datasets vary significantly - all individually are < 0.5TB.

1.4.4 Crop Wild Relatives and Genetic Resources for Food Security

Climate, occurrence, and soil datasets are used. Occurrence data complies to DwC. Climate/Soil data provides essentially raw data comprises 19 bioclimate variables combined in one file following a controlled vocabulary (https://www.worldclim.org/data/bioclim.html). However, the vocabulary for bioclimatic variables doesn't comply to the FAIR principles (e.g. I2 - https://www.go-fair.org/fair-principles/i2-metadata-use-vocabularies-follow-fair-principles). Modeller will provide mappings of bioclimate variable terms to FAIR vocabularies with WP5 team (e.g. SWEET - http://sweetontology.net/sweetAll). GBIF data is downloaded for specific taxa via API, 3GB max. Climate datasets for forcing variables are 5GB max.

1.4.5 Prioritizing Future eDNA Sampling

The pDT may use any or all (eDNA) datasets available for the relevant taxonomic group, available on GBIF. GBIF can be considered on big dataset that has ne DOI. Due to the GBIF citation tracking system, any GBIF dataset (full or filtered) will have links to the contributing datasets, which will have each their own description that can be read on GBIF.org (and available as an EML file in the original archive.) The creator and "owner" of each of the 90.000+ datasets mediated to GBIF can add metadata to their dataset. Issues raised on the GBIF mediated version will be mediated to the data owner for curation. The largest single dataset (Global Soil Organisms, https://www.gbif.org/occurrence/download?dataset_key=9f0e1ca6-fb08-4c72-9a4a-1e3b7a528c10) is 4-12 GB depending on file format. If this is converted to a site/species matrix (OTU table) it will likely take up more space. Eventually, if all available datasets of eDNA for a given taxonomic group is downloaded from GBIF, the files will be many 100 GBs (with growing available data in GBIF).

1.4.6 Phylogenetic Diversity

The model (PhyloNext) may use the full data or any subset (corresponding to e.g. to all occurrences of mammals) of the compound GBIF mediated and standardized data. GBIF can be considered on big dataset that has ne DOI. Due to the GBIF citation tracking system, any GBIF dataset (full or filtered) will have links to the contributing datasets, which will have each their own description that can be read on GBIF.org (and available as an EML file in the original archive). oToL is one big synthetic file, but the model may also just use a subset of the full tree (dataset) corresponding to e.g. the part of the tree with all mammals. The creator and "owner" of each of the 90.000+ datasets mediated to GBIF can add metadata to their dataset. Issues raised on the GBIF mediated version will be mediated to the data owner for curation. The full GBIF occurrence dataset snapshot is around 200 GB. The full synthetic tree of oToL (if downloaded, the model is not doing that) is around 28 MB.

1.4.7 Invasive Alien Plant Species

For invasives pDT, datasets are described, however, not always in a machine-readable way. We plan to create RO-Crates for each workflow run with descriptions of the input/output data. The OPenDAP server will also automatically generate data descriptors using the standard Data Access Protocol. Taimur Khan and Ahmed el-Gabbas in the project can add metadata descriptions as needed. The data sizes on the 1st workflow run: ~500GB, 2nd workflow onwards: ~50-60GB, for each run.

1.4.8 Disease Outbreaks

No current information is available at the time of the report.

1.4.9 Pollinators

From the input data described above input files are computed to run the model. The number of input files varies from two up to several thousands. Metadata needs to be developed for these files. The pDT team is in charge of updates while the project runs. Depending on the application, data files vary from several megabytes (local) to several gigabytes (country level).

1.5 Data Structure

Only a fraction of the pDT teams report the use of particular data standards or well-documented data models. However, file formats are commonly reported.

1.5.1 Grassland Biodiversity Dynamics

Original source data files are netCDF- and tiff-files. The prepared model input data file structures, contents and examples are described in a public guideline at https://zenodo.org/records/10125790. The model input data are provided to the model as .txt files (as described in the guideline).

1.5.2 Forest/bird Biodiversity Dynamics

The input files to run LANDIS-II and HMSC are combinations of .TIF (maps) and .TXT files.

1.5.3 Real-Time Bird Monitoring with Citizen Science Data

A typical data file is an audio recording (.wav files) containing sounds from 5 seconds to 60 mins. These will be used in combination with weather, land cover, and bird data. When necessary additional readme files are provided.

1.5.4 Cultural Ecosystem Services

Data details are only provided through UKCEH data management plan.

1.5.5 Crop Wild Relatives and Genetic Resources for Food Security

Typical data files are raster data in GeoTiff format and occurrence data in CSV format compliant to DwC-A. Data format is standardised as CSV for GBIF and GeoTiff for climate and soil. DwC doesn't enforce a controlled vocabulary for its term yet. Some data providers use FAIR-compliant values, some others us free text. For example, Pangaea (Involving ENVO etc): 14_4|2012-10-09T20:29:38|Pangaea|doi:10.1594/PANGAEA.299517|HumanObservation|7755606_14|Zellers, Sarah D|1992-10-05T09:35:00|Oridorsalis umbonatus|Foraminifera|Chromista|WGS84|48.6993|-126.868|R|relative abundance|-1322.0|http://purl.obolibrary.org/obo/ENVO_00002007.

1.5.6 Prioritizing Future eDNA Sampling

The occurrence data from GBIF is a simple tabular text file (with fields from the Darwin Core Standard), where each row corresponds to an occurrence. It may need to be converted into matrix format, depending on how the model is being designed.

1.5.7 Phylogenetic Diversity

The occurrence data from GBIF is a simple tabular text file (with fields from the Darwin Core Standard), where each row corresponds to an occurrence. The tree file from oToL is a text file in the Newick format.

1.5.8 Invasive Alien Plant Species

Data formats planned to be used: NetCDF4, HDF5, CSVs (maybe GeoTIFFs).

1.5.9 Disease Outbreaks

Occurrence and infection data: csv files. Landscape data: raster layer. The metadata and documentation are work in progress.

1.5.10 Pollinators

The input files of the BEEHAVE model are txt files (see example files distributed within the BEEHAVE model https://beehave-model.net).

1.6 Data Updates

1.6.1 Grassland Biodiversity Dynamics

Input data will be updated on demand, either when they are needed for a new location and/or new simulation time period or when source datasets have been updated or snapshots made available for another time point. Versioning of scripts/building blocks for retrieving and processing input data are on https://github.com/BioDT. Documentation of when which source data were used to generate the data input files is planned.

1.6.2 Forest/bird Biodiversity Dynamics

Datasets are expected to be updated but the frequencies differ between datasets. For example, LUKE updates NFI data every two years. LUOMUS updates birds' data every year. Versioning is not yet sufficiently documented and transparent, but in the future, there are plans to include versioning in the documentation. This will detail when and which source data sets are used in preparing the input files for the models.

1.6.3 Real-Time Bird Monitoring with Citizen Science Data

Dataset refresh depends on the dataset (e.g., LUKE updates NFI data every two years causing information to be more static than bird data where line transects are updated on annual basis, but LIFEPLAN data is more real-time. Partially, updates are well-recorded, but there is no practical solution for updating the version update pace in an integrated manner.

1.6.4 Cultural Ecosystem Services

Update information is only available through UKCEH data management plan. Versioning is document for some of the ~122 datasets at UKCEH.

1.6.5 Crop Wild Relatives and Genetic Resources for Food Security

Dataset updates https://www.gbif.org/occurrence-snapshots: Monthly- ERA5: Monthly with a delay of 2-3 months relatively to the actual date. Versioning will be done with folder and file names for each workflow run.

1.6.6 Prioritizing Future eDNA Sampling

GBIF occurrence data is growing day-by-day (https://www.gbif.org/analytics/global), so when/if occurrence data is retrieved from the APIs (or from program using the APIs, or the GUI / website) there will be daily differences. GBIF data does not have versions, as such. Instead, DOIs are produced that refer to GBIF datasets – either the whole GBIF dataset, or derived datasets, or single datasets. DOIs of derived datasets are linked to the contributing datasets of a filtered subset of GBIF.

1.6.7 Phylogenetic Diversity

GBIF occurrence data is growing day by day (https://www.gbif.org/analytics/global), so when/if occurrence data is utilized from the APIs (or from program using the APIs, or the GUI / website) there will be daily differences, but the snapshots (if that approach is used) are only monthly. oToL is producing a new synthetic tree on an approximately half to yearly basis: https://tree.opentreeoflife.org/about/progress?highlight=2023-09-25. GBIF data does not as such have versions. But DOIs are produced that refer to GBIF datasets – either the whole GBIF dataset, or derived datasets, or single datasets. DOIs of derived datasets are linked to the contributing datasets of a filtered subset of GBIF.

1.6.8 Invasive Alien Plant Species

The workflows only keep the newest data and not the entire dataset again. Versioning will be done with folder and file names for each workflow run, described by an RO-Crate (JSON-LD) at each run in each folder.

1.6.9 Disease Outbreaks

No updates for the current prototype version. For future model iterations updates will likely be manual for the foreseeable future. Versioning needs to be coded into the pDT.

1.6.10 Pollinators

Weather data is updated externally every day (DWD), landcover data potentially updated each year, calibration data could also update each year. Versioning is under development.

1.7 Data Sharing

1.7.1 Grassland Biodiversity Dynamics

Source data is already publicly available (e.g. Copernicus, GBIF taxonomic backbone). Currently not aware of conflicts/challenges. Identifiers of the scripts/building blocks are on https://github.com/BioDT. APIs from some of the public data sources used (e.g. https://cds.climate.copernicus.eu/api-how-to, https://pygbif.readthedocs.io/en/latest/, https://pypi.org/project/deims/, https://rest.isric.org/soilgrids/v2.0/docs).

1.7.2 Forest/bird Biodiversity Dynamics

Not aware of any conflicts at the moment. The (processed) input and generated output files will be made publicly available. Licensing is under evaluation. Currently, persistent identifiers are not used. There are no APIs for the processed and prepared input files yet.

1.7.3 Real-Time Bird Monitoring with Citizen Science Data

We are not aware of any conflicts at the moment. The (processed) input and generated output files will be made publicly available in due course. Unaware of any issues relating to licensing. Making data findable involves enhancing its accessibility and discoverability through dataset descriptions, metadata, and the use of persistent and unique identifiers (e.g., DOIs). Metadata serves as a comprehensive description of the dataset, including information such as title, author, keywords, methodology, and usage rights. Persistent and unique identifiers (DOIs) are used to ensure data discoverability (supported by University of Jyväskylä Digital Repository). Metadata contributes to the dataset's findability by enabling accurate search, retrieval, and navigation within repositories or databases. There are no APIs for the processed and prepared input files yet.

1.7.4 Cultural Ecosystem Services

Input data will be openly available but personal scores delivered by the DT user will be confidential – pDT teams are currently working out on the details. No licensing issues are known. No information on persistent identifiers available. Some API access is available.

1.7.5 Crop Wild Relatives and Genetic Resources for Food Security

There are no known conflicts or challenges in making some of the data open publicly. Licences: GBIF: https://www.gbif.org/terms, ERA5: EUMETSAT License https://cds.climate.copernicus.eu/api/v2/terms/static/licence-to-use-copernicus-products.pdf have to investigate if this corresponds to https://creativecommons.org/licenses/by/4.0/. Prefer to preserve original PID. Provider API enlisted above, CWR data will be mobilized for DEDL which provides data via HDA API (https://hda.data.destination-earth.eu/docs)

1.7.6 Prioritizing Future eDNA Sampling

The model is only using GBIF datasets that are shared under CC0, CC BY, CC BY-NC. All GBIF datasets are provided with a DOI. For GBIF occurrences (as above) there are DOIs for eDNA relevant subsets of GBIF, in which case it is called a derived dataset. DOIs of derived datasets refer to the contributing datasets. API is available but accessing eDNA data specifically (only) is currently not possible.

1.7.7 Phylogenetic Diversity

The model is only using GBIF datasets are shared under CC0, CC BY, CC BY-NC licenses, Open Tree of Life – under CC0. All GBIF datasets are provided with a DOI. For GBIF occurrences (as above) there are DOIs at the full GBIF data level (and the snapshots), but also for single datasets, and for subsets of GBIF, in which case it is called a derived dataset. DOIs of derived datasets refer to the contributing datasets. The oToL versions have versioning and are stored in an accessible place. APIs are available.

1.7.8 Invasive Alien Plant Species

All (processed) input and output data will be publicly available through the LUMI Object storage's public interface. The datasets will also be available with the OPeNDAP server running separately. Licensing requirements for eLTER sites are not completely clear and this is currently under discussion. Persistent identifiers are being explored but nothing concrete has been set at this stage.

1.7.9 Disease Outbreaks

One of the challenges in making some of the data open publicly is in sharing precise information about infected animals. We will need to investigate the issue further with our collaborators. Persistent identifiers and APIs not available at the moment of reporting.

1.7.10 Pollinators

Among the challenges in making some of the data open publicly - explicit spatial locations of beehives must not be made public. So far, only informal agreements with data owners have been made. No licenses or persistent identifiers available, only weather data API (rdwd).

Appendix I

Shared data sources for BioDT

This page lists the common data requirements across pDTs. For data sources that are commonly consumed among models of pDT, we plan to create unary data stores.

Climate Data (historical + predicted)

pDT name	variables	temporal resolution	spatial coverage	source
Grassland biodiversity dynamics	*same as in the Table Weather Data (historical + predicted) below, but as future scenarios*			
Forest/bird biodiversity dynamics	max. temperature, min. temperature, photosynthetically active radiation, precipitation, CO2 concentration	monthly - we averaged the monthly data from the last 30 years and projections of future monthly changes according to two different radiative forcing scenarios, namely RCP 4.5 and RCP 8.5.		Earth System Grid Federation (for now), DestinE (maybe later)
Pollinators	max. daily temperature, daily	daily	km^2	DWD, Copernicus (for now), DestinE (maybe later)

	sunshine hours, (maybe PAR) optional and per optimal: daily rainfall hours			
Invasive species	all Bioclim variables + PET (if available)	long-term average	spatial resolution 1 km2, Europe	CHELSA
Cultural ecosystem services	Bioclim variables	long term average	Scotland Ideally 100m resolution	https://developers.google.com/earth-engine/datasets/catalog/WORLDCLIM_V1_BIO
Crop wild relatives and genetic resources for food security	all Bioclim variables	long term average	spatial resolution 1 km2, for tropical and subtropical part of the globe	CHELSA or WorldClim (http://www.worldclim.com/version2)

Weather Data (historical + predicted)

pDT name	variables	temporal resolution	spatial coverage	source
Grassland biodiversity dynamics	**required:** air temperature [C°], precipitation [mm], solar radiation [W/m²]	daily	grassland site (no spatial variation within site)	pilot study: obtained from universities/research institutes managing sites or local weather stations, or German Weather Service (DWD, for RCP scenario predictions) general pipeline: - Copernicus ERA5-Land dataset, https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview

pDT name	variables	temporal resolution	spatial coverage	source
	calculated (i.e. part of datastream preparation)**:** potential evapotranspiration PET [mm], photosynthetic photon flux density PPFD [$\mu mol_{photons}/m^2/s$], day length DL [h]	daily	grassland site (no spatial variation within site)	PPFD calculated based on solar radiation, DL obtained based on-site latitude, PET calculated based on available weather data (cf. above, additional data stream from Copernicus API for this purpose)
	optional: rel. air humidity []	daily	grassland site (no spatial variation within site)	
Pollinators	see Table Climate Data (historical + predicted) above			

Land Use & Land Cover Data (historical + predicted)

pDT name	land use classes	temporal resolution	spatial coverage	source
Grassland biodiversity dynamics	mowing (cut grass to a specified height)	daily (mowing days are provided)	grassland site (no spatial variation within site)	pilot study: obtained from universities/research institutes managing sites, plus virtual scenarios created by modeler general pipeline: - default assumptions representing typical land use - Copernicus HRL Grassland raster: https://land.copernicus.eu/en/products/high-resolution-layer-grassland?tab=roadmap

pDT name	land use classes	temporal resolution	spatial coverage	source
				- Lange et al. 2022: https://doi.org/10.1016/j.rse.2022.112888 - Schwieder et al. 2022: https://doi.org/10.1016/j.rse.2021.112795 - Grassland mowing detection tool (for 2017-2022), https://ec.europa.eu/enrd/evaluation/knowledge-bank/grassland-mowing-detection-tool_en.html
	fertilization (add nitrogen to soil)	daily (fertilization days and amounts provided)		pilot study: obtained from universities/research institutes managing sites general pipeline: - default assumptions representing typical land use - Lange et al. 2022: https://doi.org/10.1016/j.rse.2022.112888
	land cover type (to check if a given site is 'grassland')	annual (mostly available just for one specific year)	grassland site (no spatial variation within site)	general pipeline (options): - Preidl et al. 2020 (for 2016): https://doi.org/10.1594/PANGAEA.910837 - Pflugmacher et al. 2018 (for 2015): https://doi.org/10.1594/PANGAEA.896282 - Copernicus HRL Grassland raster, Europe 2020 (for 2015, 2017, 2018, 2019, 2020 ...): https://doi.org/10.2909/60639d5b-9164-4135-ae93-fb4132bb6d83, - Eunis EEA habitat types (only for DEIMS sites): https://eunis.eea.europa.eu/habitats-code-browser.jsp?expand=290,86,1743,2421,2891,525#level_525, - Copernicus CORINE Land Cover (for ... 2006, 2018, 2024 ...): https://land.copernicus.eu/en/products/corine-land-cover,

pDT name	land use classes	temporal resolution	spatial coverage	source
				- ESA World Cover (for 2020, 2021 ...): https://zenodo.org/records/7254221
Pollinators	land cover (e.g. oilseed rape, grassland,....)	annual (phenology, i.e. flowering time on a daily basis)	Germany	Preidl et al. 2020: https://doi.org/10.1594/PANGAEA.910837
Invasive species	land cover/habitat composition	annual or average across years	Europe	Copernicus/CORINE, Copernicus
	road density	annual or average across years	Europe	Global Roads Inventory Dataset (GLOBIO)
Cultural ecosystem services	land cover/habitat composition	long term average	Scotland Ideally 100m resolution	https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_DYNAMICWORLD_V1 https://www.ceh.ac.uk/data/ukceh-land-cover-maps
Genetically detected biodiversity, cryptic habitats	land use / land cover	present state	Denmark (sine resolution)	Levin, G. 2022. Basemap04. Documentation of the data and method for the elaboration of a land use and land cover map for Denmark. Aarhus University, DCE – Danish Centre for Environment and Energy, 77 pp. Technical Report No. 252 (map data downloadable here: https://envs.au.dk/en/research-areas/society-environment-and-resources/land-use-and-gis/basemap/basemap04-geotiff-for-download)

Soil Data

pDT name	variables	temporal resolution	spatial coverage	source
Grassland biodiversity dynamic	soil composition / particle size distribution: silt content, clay content, sand content	none	grassland site (no spatial variation within site)	pilot study: local site measurements supported by literature and composition library: https://web.archive.org/web/20190405121414/https://www.nibis.de/~trianet/soil/boden4.htm general pipeline: soil maps of different horizontal and vertical resolution/extent • SoilGrid250: Europe, 250m, ○ soil layers: 0-5cm, 5-15cm, 15-30cm, 30-60cm, 60-100cm, 100-200cm ○ https://soilgrids.org/ ○ API: https://rest.isric.org/soilgrids/v2.0/docs • Ließ-map: Germany, 100m, ○ soil layers: 0-200 cm in 20 cm layer thickness ○ https://www.frontiersin.org/articles/10.3389/fenvs.2021.692959/full ○ https://osf.io/gqbmd/files/osfstorage
	soil layer characteristics: field capacity, permanent wilting point, porosity, hydraulic conductivity	none	vertical layers (no lateral variation), e.g. 20 layers, each with depth 10 cm	pilot study: local measurements per layer, or calculated based on soil texture/composition using literature (i.e. Maidment Book of Hydrology) or pedotransfer functions (PTFs)

pDT name	variables	temporal resolution	spatial coverage	source

				general pipeline: soil maps if different horizontal and vertical resolution/extent
				• HiHydroSoil: based on SoilGrid250 using pedotransfer functions:
				  ○ https://www.futurewater.eu/projects/hihydrosoil
				  ○ documentation: https://www.futurewater.nl/wp-content/uploads/2020/10/HiHydroSoil-v2.0-High-Resolution-Soil-Maps-of-Global-Hydraulic-Properties_v2.pdf
				• ESDAC: Europe, 500m,
				  ○ soil layers: 0-20cm
				  ○ https://esdac.jrc.ec.europa.eu/content/maps-indicators-soil-hydraulic-properties-europe#tabs-0-description=0
				  ○ uses PTFs to derive hydrological properties based on R package: euptf (https://esdac.jrc.ec.europa.eu/public_path/shared_folder/dataset/39_hydraulic_properties/euptf_vignette.pdf)

Forest/bird biodiversity dynamic	soil type	none	Finland	Combined map of different sources: Corine Land Use (https://land.copernicus.eu/pan-european/corine-land-cover) and Soil map

pDT name	variables	temporal resolution	spatial coverage	source
				(https://www.sciencedirect.com/science/article/pii/S0264837711000718)
Invasive species	soil type, bedrock, pH	none	Europe	SoilGrids250m as one option: https://soil-modeling.org/resources-links/data-portal/soilgrids250m
Cultural ecosystem services	soil type, bedrock, pH	none	Scotland	https://developers.google.com/earth-engine/datasets/catalog/OpenLandMap_SOL_SOL_ORGANIC-CARBON_USDA-6A1C_M_v02 https://developers.google.com/earth-engine/datasets/catalog/OpenLandMap_SOL_SOL_PH-H2O_USDA-4C1A2A_M_v02
Crop wild relatives and genetic resources for food security	soil type, soil pH and mineral contents such as soil zinc content	none	Will sample to 1 km2; global scale	https://soilgrids.org/, Soil layer - 0 -30 cm

Observations Data

pDT name	variables	temporal resolution	spatial coverage	source
Grassland biodiversity dynamics	biomass (per PFT), vegetation ground cover (per PFT), (optional: max. vegetation height, leaf area index (per PFT))	pilot study/local site observations: measured at specific dates (few times per year)	pilot study/local site observations: patch(es) of 1 m² general pipeline: extent/resolution of available maps	pilot study: obtained from universities/research institutes managing local grassland sites local site observations: eLTER general pipeline: remote sensing products - Copernicus ERA5-Land dataset (e.g. low vegetation cover, LAI),

pDT name	variables	temporal resolution	spatial coverage	source
		general pipeline: e.g. daily		https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview - Copernicus HRL Grassland raster (e.g. herbaceous cover), https://land.copernicus.eu/en/products/high-resolution-layer-grassland?tab=roadmap
Forest/bird biodiversity dynamics	species occurrences	yearly	Finland	Finnish Museum of Natural History (LUOMUS)
Crop wild relatives	species occurrences (GBIF) crop trait observations, including on to what degree a grass pea genotype/variety/cultivar is causing Lathyrism, and crop pests and diseases (diseases harming the crop)	not applicable	not applicable	EURISCO, GeneSys, etc. (wherever is available including scientific literature search) - can be published in GBIF pending an upgrade of the data model
Pollinators	weight of the beehives as a proxy for honey (kg)	daily	~400 scales in Germany	https://www.dlr.rlp.de
Invasive species	species occurrences	aggregated across all years (for now)	Europe	GBIF, eLTER
Cultural ecosystem services	species occurrences	unaggregated	Scotland	GBIF, eLTER

pDT name	variables	temporal resolution	spatial coverage	source
Genetically detected biodiversity, cryptic habitats	species occurrences	all available data (withing the geographical and taxonomic scope)	Denmark	GBIF API description: https://www.gbif.org/developer/summary
Crop wild relatives and genetic resources for food security	species occurrences, trait data	not applicable	tropical and subtropical parts of the globe	GBIF, ICARDA, Genesys and crop trust

Various Other Data Sources

data	data type	data format	spatial extent	spatial resolution	temporal scope	temporal resolution	API	ex. mod use	data source/location	status
D1: GBIF occurrences	occurrences	vector (points)	global	various	1970–2000 (extended to 2023, updated regularly)	dynamic	y	model calibration	GBIF	RI (download access)
D2: Ecosystem types of Europe v3.1	probability of EUNIS habitat type presence	categorical	regional (Europe, EEA39)	100 m	2012	static	n		European Environmental Agency: https://www.eea.europa.eu/data-and-maps/data/ecosystem-types-of-europe-1	RI (download access)

data	data type	data format	spatial extent	spatial resolution	temporal scope	temporal resolution	API	ex. mod use	data source/location	status
D3: European checklist of alien plants	plant species' floristic status	categorical	regional (Europe)	country	2022	static	n		Unpublished yet, provided directly by Irena Axmanová	local storage
D4: Maps for 7 ecosystem services	ecosystem services (relative pressure by IAS on terrestrial and freshwater ecosystems, estimated for 23 plant and 26 animal species)	continuous (polygons)	regional (EU28)	10 km	2012–2017	static	n		European Commission, Joint Research Centre: https://data.jrc.ec.europa.eu/dataset/f11838ff-8984-4284-bea5-e0a11d9f8d2f	RI (download access)
D5: CHELSA climatologies	climate	continuous / categorical	global	30 rc-sec	Climatologies data: current (1981–2010), future: 3 times (2011–2040/2041–2070/2071–2100) x 5 models		n		https://chelsa-climate.org/	RI (download access)

data	data type	data format	spatial extent	spatial resolution	temporal scope	temporal resolution	API	ex. mod use	data source/location	status
					(CMIP6) x 3 scenarios					
D6: Global Roads Inventory Dataset [GRIP]	road lines (ESRI filegeo database / Shapefile), road density (ASCII)	continuous	global	density: 5 arc-min (~8x8km)	2018	static	n		https://datacatalog.worldbank.org/search/dataset/0037825 GRIP global roads database: https://www.globio.info/download-grip-dataset	RI (download access)
D7: ESA WorldCover	land cover/land use	categorical	global	10 m	2020, 2021	static	?		https://esa-worldcover.org/en	RI (download access)
D8: Pan-European High-Resolution Layers	land cover/land use		Europe	10 m, 20 m (depending on year)	2006-2018	static	n		https://land.copernicus.eu/pan-european/high-resolution-layers	RI (download access)
D9: SoilGrids250m	soil characteristics	various	global	250 m	2016	static	n		https://soil-modeling.org/resources-links/data-portal/soilgrids250m	RI (download access)