BIODT
biodiversitydigitaltwin

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

D1.2 - Data Management Plan

Work Package	WP1 | Project Management
Task	T1.2 | Quality Control, Ethics, Risk Management and DMP
Lead Authors	Jesse Harrison (CSC), Anna-Liisa Allas (CSC), Hanna Koivula (CSC), Roni Blankett (CSC)
Contributors	
Peer Reviewers	Wouter Addink (Naturalis), Dag Endresen (UiO), Nicola Fiore (LifeWatch ERIC)
Version	V1.0
Due Date	30/11/2022
Submission Date	25/11/2022

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	28/10/2022	Jesse Harrison, Anna-Liisa Allas, Hanna Koivula, Roni Blankett	First draft pending feedback from peer reviewers
0.2	10/11/2022	Jesse Harrison, Anna-Liisa Allas	Version incorporating feedback from peer reviewers
0.3	25/11/2022	Jesse Harrison	Version incorporating feedback from BioDT Project Management Board (PMB) and Council
1.0	25/11/2022	Jesse Harrison, Anna-Liisa Allas	Submission-ready version incorporating BioDT Project Management Office (PMO) editorial changes

Glossary of Terms

Item	Description
API	Application Programming Interface
DCAT	Data Catalog Vocabulary
DestinE	Destination Earth
DiSSCo	Distributed System of Scientific Collections
DMP	Data Management Plan
DOI	Digital Object Identifier
DT	Digital twin
EC	European Commission
eLTER	Integrated European Long-Term Ecosystem, critical zone and socio-ecological Research
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
FDO	FAIR Digital Object
GBIF	Global Biodiversity Information Facility
GDPR	General Data Protection Regulation
IPP	Intellectual Property Policy

IPR	Intellectual Property Right
JSON	JavaScript Object Notation
LifeWatch ERIC	E-Science European Infrastructure for Biodiversity and Ecosystem Research
maDMP	Machine actionable Data Management Plan
Naturalis	Naturalis Biodiversity Center
OpenAIRE	Open Access Infrastructure for Research in Europe
PID	Persistent identifier
RDF	Resource Description Framework
RI	Research Infrastructure
UC	Use Case
UiO	University of Oslo
W3C	World Wide Web Consortium

Keywords

Biodiversity, Data management, Digital Twin, FAIR data, Intellectual property, Modelling, Sustainability

Executive Summary

This document describes the Data Management Plan (DMP) for data collected, processed and created by the BioDT project. It presents a description of data lifecycle management within the project, from data reuse or data generation to storage, preservation and further use by third parties while applying FAIR (Findable, Accessible, Interoperable and Reusable) data principles. The DMP also incorporates information on recommended licensing schemes to be used, formulated as an intellectual property policy for the BioDT project.

The BioDT DMP is a living document that will be updated, reviewed and versioned to incorporate changes and additions throughout the lifetime of the project.

Table of Contents

List of Tables

1. Introduction

The Biodiversity Digital Twin (BioDT) project aims to push the current boundaries of our understanding of biodiversity dynamics by developing Biodiversity Digital Twin (BioDT) prototypes that provide advanced modelling, simulation and prediction capabilities. The project will exploit existing technologies and data available across relevant biodiversity research infrastructures in new ways, making it possible to accurately and quantitatively model interactions between species and their environment.

This document describes the Data Management Plan (DMP) for data collected, processed and created by the BioDT project, in line with the European Commission (EC) Open Access to research data policy for facilitating access, reuse and preservation of its research data.

The BioDT DMP is a living document that will be updated, reviewed and versioned to incorporate changes and additions throughout the lifetime of the project. Prior to their submission to the European Commission Funding & Tenders Portal, new versions of the DMP will be circulated for feedback from and approval by the BioDT Project Management Board (PMB) and the BioDT Council.

Since the BioDT DMP is a living document, any version can vary in length and level of detail following the progress of the project. In order to keep track of the updated versions of the DMP, the version number is always included on the title page of the DMP.

The DMP provides information on:

- Usage of data during and after the project, including existing data and data produced by other research or infrastructure initiatives
- Types and formats of data to be collected, processed or generated
- Data sources and anticipated data storage requirements
- The anticipated utilities of data produced by BioDT
- Methodologies and standards for data handling, including compliance with FAIR (Findable, Accessible, Interoperable, Reusable) principles
- Practices to be followed in the handling of sensitive or otherwise (e.g. legally or contractually) restricted data
- Information on recommended licensing schemes to be used, formulated as an intellectual property policy for the BioDT project

Further details to be included in subsequent versions of the DMP are outlined in Section 10.

2. Data Summary

The BioDT project will develop prototype digital twins (DT) leveraging high-performance computing, predictive biodiversity modelling methods and biodiversity data streams from operative biodiversity research data infrastructures to address eight different use cases. The BioDT Use Cases address research questions spanning several levels of complexity, which define the types, formats and volumes of data to be used. The systems to be simulated by the BioDT Use Cases may be:

- Simple (e.g. a two-species system)
- Intermediate (e.g. local-scale data on a well-studied arctic ecosystem)
- Complex (e.g. local-scale data on a well-studied tropical ecosystem), or
- Highly complex (e.g. spatially extensive data covering interactions among several ecosystem types)

Within the systems to be simulated, the complexity of biotic and abiotic data to be used may also vary. As an example, the biological datasets to be used may range from simple species presence-absence data to more complex (e.g. genetic and/or hierarchical) data, and from univariate to multivariate datasets.

The BioDT project relies on reusing biodiversity data made available via the four partnering European biodiversity research infrastructures (GBIF - the Global Biodiversity Information Facility, eLTER - the European Long-Term Ecosystem, critical zone and socio-ecological Research Infrastructure, DiSSCo - Distributed System of Scientific Collections, and LifeWatch ERIC - the e-Science European Infrastructure for Biodiversity and Ecosystem Research), and mobilizing new biodiversity data sources through these infrastructures, as well as abiotic data provided by sources including the Destination Earth (DestinE) initiative overseen by the European Commission. Through its modelling scenarios and by aggregating biotic and abiotic data from multiple sources that could be used as analysis-ready data sets by other DT projects, BioDT aims to contribute to the goal of DestinE to construct a full DT of the Earth.

This section summarises the ways in which BioDT will reuse existing data, the types, formats and expected size of data to be reused or generated, as well as the sources, anticipated storage requirements and utility of data to be used and produced by the project.

The BioDT project employs an iterative approach to Use Case development. During earlier stages of the project, Use Case and DT development will focus on limited and readily available data sets that can be handled locally. Subsequent iterations of the Use Cases and the associated DTs will be deployed on the EuroHPC LUMI supercomputer (https://www.lumi-supercomputer.eu). The types and formats of data to be reused and generated by the project, as well as data storage requirements, will vary depending on project progress and the Use Case iterations under development.

2.1 Reuse of Existing Data

Existing data sets accessed via BioDT-affiliated biodiversity research infrastructures (RIs) will be used as input data for DTs constructed to address each BioDT Use Case, and for developing the modelling approaches and other data processing or analysis methods to be used. As part of DT development and testing, the data will be used to assess the predictive performance of the DTs and to confirm the functionality of DT end-user features (e.g. for the purposes of biodiversity monitoring).

Provision of biodiversity data to the digital twins developed in BioDT will be arranged via GBIF or through other biodiversity RIs using a standard application programming interface (API), with alternative solutions established where required by individual Use Cases. For example, should a BioDT Use Case require a richer version of the original data set compared to that made available via the biodiversity RIs, the original source data may be used. Readily available data sets may also be used directly for initial prototype DT development work.

Further to data sets accessed via biodiversity RIs, the project may use existing data sets generated by DestinE, other DT projects, BioDT-affiliated research groups and their collaborators. Where required, data may also be used from RIs otherwise external to the project, including COPERNICUS and ICOS. The data to be used depend on the requirements and objectives of each BioDT Use Case.

2.2 Types and Formats of Data to be Reused or Generated

2.2.1 Definitions of Data and Data Types

The BioDT project employs a definition of data that includes data reused by the project, as well as diverse project outputs (such as code, scripts, software, workflows, models, services and end-user features).

Data reused by the project include existing third-party data products, i.e. data sets that are quality controlled and analysed or modelled, consistent through time and preferably available through an API. Examples of data products are nature type or land cover data available (as map layers) from national land surveys, or national plant or bird atlases, where data are collected with a method and harmonised to express absence data, abundance and possibly also changes through time. Data can also be raw or derived data compiled from multiple sources, which may be further refined to ensure compatibility with the BioDT Use Cases and the DTs employed by them.

Data generated by BioDT are defined to include aggregated and/or modified data sets based on the reuse of existing data (e.g. composite data sets including biotic and abiotic data), code (e.g. code used for software, workflows or different DT components), as well as other project outputs (e.g. complete DTs, new software versions, models, training and dissemination materials). Data generated by the project also include all analytical outputs that will be made openly accessible as results and to ensure reproducibility of analysis and modelling approaches used in the project.

Specific data types used by BioDT are outlined in Section 2.2.2. The application of FAIR management principles to data to be reused and generated by the project are described in Sections 3 and 4.

2.2.2 Data Types and Formats Reused and Generated by BioDT

An overview of data types and formats to be reused by the BioDT project is provided in Table 1, while data types and formats to be generated are described in Table 2.

Data Types to Be Reused	Data Formats to be Reused
Biotic data and data accompanying them, including: • Taxonomic information – E.g. taxon occurrence and abundance data, taxon checklists (defining taxonomic composition of an occurrence data set including absence data), disease vector and pathogen distribution data, EU and national lists of alien invasive or endangered species (categories in legislation) • Ecological and species interactions data – E.g. plant and pollinator quantitative and interaction web data, plant functional types • Genomic and biochemical data – E.g. DNA sequence data, specimen data • Acoustic and image / video data – E.g. bird song recordings and camera trap data	• Hierarchical data (HDF5) • Object and FAIR Digital Object data, serialised as JavaScript Object Notation (JSON) data and JSON for Linking Data (JSON-LD) • Text data (txt, csv) • Generic (provenance, rights, licenses) and domain-specific metadata using vocabularies that follow the FAIR principles like schema.org and DCAT • *Additional formats to be confirmed*
Abiotic data, including: • Geographic data – E.g. geographic location data, land use data • Nature types and habitat type classifications	• Text data (txt, csv) • Image files (e.g. GeoTIFF) • State vector data • *Additional formats to be confirmed*

- Carbon stocks - Water quality - Climate variables (e.g. precipitation and temperature) - Trade exchange and traffic - Human density data, urban footprint data	

Table 1 – Overview of data types reused by BioDT Use Cases

Data Types to Be Generated	Data Formats to Be Generated
Aggregated biotic and abiotic data sets	- Hierarchical data (HDF5) - Text data (txt, csv)
Model output data	- Text data (txt, csv) - Hierarchical data (HDF5)
- Source code for features including: – Computational workflows – Data analysis scripts – Software and modelling tools – DT components, including end-user features - Outputs based on the above (e.g. complete workflows, DTs and tools for using DTs, software)	- Text data (txt, csv) - JavaScript Object Notation (JSON) data - Data analysis scripts (e.g. Python, R, NetLogo, Rust) - Scripts used by computational workflow managers (e.g. Apache Airflow) - Workflow scripts including semantic artefacts providing rich provenance information (e.g. CWL, W3C PROV serialised as JSON-LD) - *Additional data formats to be confirmed*
Method and protocol descriptions	- Several formats (txt, csv, PDF, html)
Dissemination and communication materials	- Training materials based on several formats (e.g. PPT, PDF, html)

Table 2 – Overview of data types generated by BioDT

Key data sets based on reformatted and combined versions of the original data (to be used as DT input data or produced as model outputs) will be made available for reuse by biodiversity RIs, other DT projects, academic researchers and additional target audiences of the project, following FAIR principles (Section 3).

2.3 Data Sources

Details on data sources employed by each BioDT Use Case are provided in Table 3. **The sources listed are dependent on Use Case development and the current iteration of the Use Case.** **In addition to the sources listed in Table 3, data sources used by BioDT can include other digital twin initiatives and DestinE.**

Use Case Group	Specific Use Cases (Where Applicable)	BioDT-affiliated RIs Serving as Data Sources	Third-party Data Sources
Biodiversity Dynamics	Grassland Biodiversity Dynamics	- GBIF - eLTER	- *To be confirmed*

		- *Other RIs to be confirmed*	
	Forest / Bird Biodiversity Dynamics	- GBIF - *Other RIs to be confirmed*	- Natural Resources Institute Finland (https://www.luke.fi/en) - Finnish Museum of Natural History (https://luomus.fi/en) - Earth System Grid Federation (https://esgf.llnl.gov) - *Other sources to be confirmed*
	Real-time Bird Monitoring Using Citizen Science Data	- GBIF - *Other RIs to be confirmed*	- Finnish Biodiversity Information Facility (https://laji.fi/en) - *Other sources to be confirmed*
Ecosystem Services	Cultural Ecosystem Services	- GBIF - *Other RIs to be confirmed*	- *To be confirmed*
Crop Wild Relatives, Genetic Resources for Food Security		- GBIF - *Other RIs to be confirmed*	- *To be confirmed*
DNA Detected Biodiversity, Poorly Known Habitats		- GBIF - *Other RIs to be confirmed*	- Global Register of Introduced and Invasive Species (https://griis.org) - *Other sources to be confirmed*
Endangered Species		- GBIF - *Other RIs to be confirmed*	- *To be confirmed*
Invasive Species	Alien Plant Species Dynamics	- GBIF - *Other RIs to be confirmed*	- *To be confirmed*
Disease Outbreaks	Disease Vector Dynamics	- GBIF	- *To be confirmed*

		• *Other RIs to be confirmed*	
	Disease Transmission Between Wild and Domesticated Populations	• GBIF • *Other RIs to be confirmed*	• ENETWILD (https://enetwild.com) • *Other sources to be confirmed*
Pollinators	Honey Bee Dynamics in Agricultural Landscapes	• GBIF • *Other RIs to be confirmed*	• *To be confirmed*

Table 3 – Data sources utilised by each BioDT Use Case

2.4 Expected Storage Requirements

Data storage requirements of each BioDT Use Case are described in Table 4. Because BioDT relies on dynamic data streams, data sizes specified in this Data Management Plan are described in terms of anticipated storage requirements. **The data storage requirements of each Use Case vary depending on the development status of each Use Case and represent the current maximum in terms of the expected data volume.**

Use Case Group	Specific Use Cases (Where Applicable)	Storage Requirement(s) for Input and Output Data
Biodiversity Dynamics	Grassland Biodiversity Dynamics	Several gigabytes (Gb)
	Forest / Bird Biodiversity Dynamics	Several gigabytes (Gb)
	Real-time Bird Monitoring with Citizen Science Data	Several terabytes (Tb)
Ecosystem Services	Cultural Ecosystem Services	Several gigabytes (Gb)
Crop Wild Relatives, Genetic Resources for Food Security		Several gigabytes (Gb)
DNA Detected Biodiversity, Poorly Known Habitats		Several gigabytes (Gb)
Endangered Species		Several gigabytes (Gb)
Invasive Species	Alien Plant Species Dynamics	Several gigabytes (Gb)
Disease Outbreaks	Disease Vector Dynamics	Several gigabytes (Gb)
	Disease Transmission Between Wild and Domesticated Populations	Several gigabytes (Gb)

Pollinators	Honey Bee Dynamics in Agricultural Landscapes	Several gigabytes (Gb)

Table 4 – Expected input and output data storage requirements

2.5 Data Utility

The BioDT project relies on data-driven predictive simulation and modelling tasks, the results of which are expected to be useful to researchers working in the fields of biodiversity research, conservation biology, ecology and the environmental sciences. In particular, scientists will be able to better observe changes in biodiversity in response to forces resulting from climate change or human activity, mechanistically understand how these changes occur, and predict the effects of these changes. The biodiversity RI nodes and researchers from different disciplines can also use the results of BioDT to improve their services.

Aggregated data sets produced by BioDT also have the capacity to contribute toward other digital twinning projects, including those implemented as part of DestinE. The prototype DTs constructed during the project hold potential in serving as a blueprint for the development of increasingly powerful digital twinning solutions that can be used by policymakers to better respond to societal needs and key initiatives, including the EU Biodiversity Strategy 2030 and Sustainable Development Goals established by the United Nations.

Furthermore, industrial actors and small to medium-scale enterprises may be able to exploit BioDT for business solutions in sectors related to biodiversity, such as agri-food, tourism and healthcare. The BioDT project will also be useful for boosting citizen science, strengthening common understanding of biodiversity dynamics and prediction models, and fostering biodiversity literacy and trust in biodiversity research among the general public.

3. FAIR Data

Building digital twins to study global biodiversity dynamics requires data from a cross-domain perspective and data derived from a wide spectrum of sources (see Section 2). Over the years, FAIR principles have emerged as a guideline for addressing such heterogeneous and complex data sources. We use FAIR principles as a reference point to ensure that data within the BioDT platform can be:

- Organised and managed to be easily *Findable*
- *Accessible* as widely and easily as possible, in a manner that is user-friendly, machine-readable, and machine-actionable
- *Interoperable* so it can be linked with other data and Digital Twins, and
- *Reusable* and *reproducible* so it is easy to re-purpose and re-create.

FAIR principles will be applied in BioDT by taking the following actions. Where limitations or restrictions to data openness apply, steps for the handling of such data are outlined in Section 3.2.

3.1 Making Data Findable, Including Provisions for Metadata

Data sets needed for the models and applications to be used and developed by BioDT require new combinations of existing biodiversity and abiotic data (Section 2.2), accompanied by processes to ensure transparency, credit and attribution in relation to the original source data. To achieve this, the BioDT project will use established norms for producing aggregated biodiversity data sets where possible. For example, data downloads from GBIF (www.gbif.org) provide digital object identifiers (DOIs) for downloads that point to the DOIs of different original data sources.

For aggregated data sets including source data without data set DOIs, data papers (see Section 3.2.1) will provide additional data citation solutions.

The usage of persistent identifiers (PIDs), typed objects and machine actionable metadata as components of FAIR Digital Objects (FDO) is designed to improve machine actionability, linkability as well as findability of the data in alignment with the European Open Science Cloud (EOSC) PID Policy and the EOSC Interoperability Framework. The FDO framework will facilitate the implementation of an object-oriented approach that is modular, actionable, meaningful, technology independent and FAIR by design. A core objective of FDOs and FAIR is "machine actionability", the capability of machines to handle data autonomously and appropriately. Using this approach, the goal is to ultimately use this DMP document to inform the development of an automatic and self-updating FDO type which enables integration and automation of all relevant processing steps in the data life cycle, implementing in this way the concept of a machine actionable data management plan (maDMP).

Following the BioDT Grant Agreement Annex 5 Article 17, metadata of deposited data must be open under a Creative Common Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded), in line with FAIR principles. The metadata must be machine-readable and must provide information at least about the following:

- Data sets (description, date of deposit, author(s), venue and embargo)
- Horizon Europe or Euratom funding
- Grant project name, acronym and number
- Licensing terms
- Persistent identifiers for the data sets
- The authors involved in the action identified by their ORCID ID and, if possible, their organisation identified by their Research Organization Registry (ROR) ID
- Where applicable, the metadata must include persistent identifiers, such as the CrossRef DOIs, for related publications and other research outputs.

Metadata standards (as well as community standards) to be used by BioDT are outlined in Section 3.3.1.

3.2 Making Data Accessible

In addition to improving Findability, the usage of PIDs, typed objects and machine actionable metadata as part of FAIR Digital Objects will improve Accessibility of data handled by the BioDT project.

3.2.1 Deposition of Data in Repositories and Data Lakes

Key data produced during the project lifetime (e.g. data sets and accompanying peer-reviewed publications) will be available via a trusted, FAIR data repository or repositories, accessible via open protocols. FAIR management principles will also be applied to project outputs including source code, scripts, computational workflow code, protocols and software (details in Section 4.1).

Data in relation to training materials produced for the biodiversity RIs and the wider research community will either be accessible via the RIs or separately deposited in open repositories for improved Accessibility. A goal of the training materials produced for the project will be to improve Accessibility by detailing how to use the biodiversity digital twin prototypes constructed as part of the project.

Data papers are peer-reviewed static data sets which are published accompanying rich metadata and citable with an identifier (for example, DataCite DOI). They will be used to publish data sets generated by the project in case they cannot be published using platforms provided by participating RIs. Data papers can consist of a wide combination of compiled data and the platform should publish the data in a structured format that allows data citation and ingestion of the data by BioDT workflows, preferably through an API. The BioDT project will maintain internal documentation on platforms and/or data journals that are suitable for publishing project outputs.

As part of BioDT, provisions will be made to grant researchers with access to specific data sets through the DestinE interface (e.g. via the data lake service in DestinE). Details on this topic are to be provided in a subsequent version of the BioDT DMP.

3.2.2 Sensitive or Restricted Data

Selected biodiversity data may have legal and/or contractual restrictions on openness, e.g. location data for endangered species or data on indigenous lands. To accommodate for use cases involving the handling of sensitive or otherwise restricted biodiversity data, BioDT will follow an "open as possible, closed as (legally) necessary" policy in relation to data set maintenance and publication.

Access to sensitive data during the lifetime of the project will follow existing procedures established by the biodiversity research infrastructures involved in the project. Therefore, establishing a separate data access committee is not required.

Where restrictions to data openness are specified in the BioDT Consortium Agreement and Grant Agreement, these prevail over the general terms outlined in the Intellectual Property Policy provided as part of this Data Management Plan (see Section 8).

3.3 Making Data Interoperable

Interoperable data means it can be integrated with other data, applications and workflows. This is achieved by using common metadata and data standards, and harmonising data by using semantic artefacts (i.e. controlled vocabularies, ontologies and thesauri, etc.) to describe the data variables unambiguously. Alignment between different data sources can be created with crosswalks between the models. Likewise, technical interoperability is achieved by creating automated workflows using standards and by using technical standards with APIs for data transfer.

3.3.1 Data, Metadata and Community Standards

The project will create FAIR metadata for workflows and models, as well as their packaging, interlinking and lifecycle. All digital twin components (including data and models) will be linked, creating support for model ensembles and the modelling lifecycle, and enable comparison, exchange, archiving and publishing with persistent identifiers. The project will deal with several data standards related to different disciplines. One of the objectives of the project is to address semantic linking and interoperability issues for different data standards.

Existing community standards to be used as part of BioDT include:

- DarwinCore and its extensions (http://www.tdwg.org/standards/450)[1]
- EML (Ecological Metadata Language)[2]
- openDS (the open Digital Specimen specification, https://github.com/DiSSCo/openDS)

The BioDT project will establish a curated standards repository that will be maintained and updated on the BioDT GitHub organisation (https://github.com/BioDT). Hosting standards in a public GitHub repository is intended to provide a transparent, easily visible and interoperable method for standard curation as part of BioDT.

Utilising Common Workflow Language (CWL)-compliant code will be used to facilitate FAIR compliance of analytical pipeline development as part of the BioDT project.

3.3.2 Semantic Mapping

Along with using FAIR principles as guidelines, a cross-domain semantic mapping will be piloted between digital objects to further improve Interoperability. As discoverability and accessibility are essential aspects of the FAIR principles, semantic mapping ensures data and other entities can be integrated and used despite heterogeneity of the baseline data sets used by BioDT. Various data staging, processing, and sharing activities in LUMI and other HPC environments can make use of semantic mapping for integrating data using different ontologies and vocabularies. The semantic mapping approach can be combined with existing tangible results produced by the European Strategy Forum for Research Infrastructures (ESFRI; https://www.esfri.eu) cluster projects, including ENVRI FAIR (https://envri.eu/home-envri-fair) and EOSC-Life (https://www.eosc-life.eu).

Each participating network harmonises their data sets according to the original purpose. BioDT will use existing vocabularies to create crosswalks to compile data sets when data are aggregated from multiple sources. Widely used existing vocabularies are, for example:

- The GBIF Backbone Taxonomy[3], which is a single, synthetic management classification with the goal of covering all names GBIF is dealing with
- EnvThes[4], the European LTER network's controlled vocabulary
- ENVO (https://sites.google.com/site/environmentontology), the environment ontology, which enables the machine-actionable knowledge representation of environmental entities and processes
- Bioschemas (https://bioschemas.org), an extension of schema.org providing new data types related to biology and life sciences (e.g. Taxon, Gene, MolecularEntity)

Variables in the metadata can be crosswalked with W3C - DCAT, which is a Resource Description Framework (RDF) vocabulary designed to facilitate interoperability between data catalogues published on the Web. However, DCAT cannot be used for data level integration.

3.3.3 Alignment with Other Projects and Initiatives

While an objective of DestinE is to create more accurate twins for e.g. climate and weather prediction, the initiative also seeks to serve the requirements of other DT projects simulating aspects of the natural environment. Whereas BioDT focuses on providing prototype DTs for simulating terrestrial biodiversity dynamics, other emerging DT projects will focus on alternative scenarios (such as ocean biodiversity dynamics as part of an Ocean Digital Twin). To help achieve the goal of DestinE to ultimately construct a full digital twin of the Earth, BioDT will need to fit into the wider landscape of DTs being established in Europe. Further to alignment with DestinE and other DT projects, the involvement of GBIF, eLTER, DiSSCo and LifeWatch ERIC as BioDT project partners ensures direct alignment with key biodiversity research infrastructures facilitating the provision and accessibility of biodiversity data.

Reaching this objective will require conceptual alignment with other twin initiatives. For example, during the iterations of developing the BioDT Use Cases, a proof of concept demonstrator(s) will be created that features meaningful interactions between twins from different initiatives. BioDT will also align with other initiatives at the level of data and metadata, and key services produced by the project will be made available in the European Open Science cloud using APIs. Another link will be made to the EU Data Spaces initiative to ensure data and the BioDT models are available for anyone in the EU, in agreement with FAIR rules and the EU principles and values for the use of data.

Tools for data aggregation, cleaning and wrangling to be used as part of BioDT will rely on established tools that are also used by other projects and initiatives (e.g. tools provided by LifeWatch ERIC, Naturalis or GBIF; see e.g. the GBIF tools catalogue at https://www.gbif.org/resource/search?contentType=tool). The use of CWL for computational workflow development (Section 3.3.1) will further improve the compatibility of BioDT project outputs with other projects, with workflows additionally deposited in the WorkflowHub FAIR workflow registry sponsored by EOSC-Life and ELIXIR (https://about.workflowhub.eu).

3.4 Increasing Data Reuse

Steps taken to increase data reuse as part of the BioDT project include using open licenses for the deposition and publication of data where possible, and establishing conceptual and technical linkages with the wider European DT landscape (see Sections 3.1-3.3). The BioDT project will also increase data reuse through training activities designed for DT end-users, the production of end-user documentation aimed at improving the transparency, repeatability and understandability of the work carried out. By adopting a broad definition of data (Section 2) and adhering to FAIR principles, the project will also seek to maximise the usability of data and project-affiliated source code by third parties. Where possible, implementing a FAIR digital object layer will also be used to introduce machine actionability, promoting data re-use by machines and automated services.

3.4.1 Documentation to Validate Data Analysis and Facilitate Data Reuse

The BioDT project will produce end-user documentation for the use of the DT platform, resulting datasets and associated DTs, an objective of which is to facilitate independent use of the data analysis tools that will be provided. To ensure transparency and repeatability, key data sets supporting scientific publications and code / scripts for reproducing the work will be made available in open repositories, complete with citation information to ensure compliance with FAIR principles.

3.4.2 Usability of Data from Third Parties

Data to be reused by BioDT (Section 2) will be provided via the project-affiliated biodiversity RIs, with these data also being accessible to third parties using standard protocols for access and data retrieval. The data provided by RIs are well defined and quality assured, and use commonly accepted (community) standards. To align the project with DestinE and other European DT initiatives, and to comply with FAIR data management principles, the project is also designed to provide open third-party access to data generated by the project where permitted by the project Consortium Agreement and/or Grant Agreement.

3.5 Assessment of FAIR Compliance

The level of FAIRness of research data output will be measured through appropriate FAIR assessment tools. These tools, along with maturity indicators, define core criteria to assess the implementation level of the FAIR data principles and provide a structured yet flexible approach in assessing FAIRness. The BioDT project will develop semi-automated indicators (e.g. usage of persistent identifiers, ensuring that metadata is represented using a formal knowledge representation language) of quality for the data sets and workflows provisioning those results. Quality indicators, such as FAIRness assessment metrics and geographic data accuracy, can provide processes for consistency that are needed in different stages of workflows using a digital twin. These indicators are also crucial for compliance with recognised community standards and benchmarking for performance tests. The level of FAIRness and other quality indicators based on the quality enhancements coming from RI data streams (e.g. geospatial and temporal data) will be used to build the quality mapping framework. Incorporating the FDO layer, workflows and data provisioning through the RIs is intended to provide researchers with reliability and verifiability of the data, while the quality mapping framework is designed to guide users about the quality of the data and expected outputs of BioDT. The BioDT Use Cases will be aligned with the FAIR framework and will provide input to develop quality indicators for the accuracy of the developed models.

4. Other Research Outputs

4.1 FAIR Management of Software, Workflows, Protocols and Models

As part of DT development activities, the BioDT project will generate new source code and data analysis code (e.g. Python, NetLogo, R and Rust scripts) to enable diverse functionalities required for the operation of prototype DTs. Such source code may pertain to the development of new software releases, workflows, protocols for data transfer, processing and analysis, the development of improved modelling methods, and the design of end-user features for interacting with DTs.

The steps described in this document in relation to adherence to FAIR principles are applicable to all data handled by the BioDT project, including services, software and workflows. Accordingly, source code and data analysis code produced by the project will be subject to handling in agreement with FAIR guidelines and will be deposited in open repositories where possible (Section 3). Handling of source code and data analysis code will adhere to commonly agreed practices established in the BioDT Consortium Agreement, the BioDT Grant Agreement and the Intellectual Property Policy outlined in this DMP (Section 8). Details on the handling of sensitive and/or restricted data are also discussed in Section 3.2.2.

Source code and scripts generated by the project will be deposited in either existing repositories (for example, GitHub repositories for specific biodiversity modelling tools) or within the BioDT GitHub organisation (https://github.com/BioDT).

Where possible, open repositories will also be used for the release of new software versions resulting from BioDT-related work. Details on software licensing are outlined in the Intellectual Property Policy included in this document (Section 8).

To further improve Findability of the different services and features developed during the project, BioDT will leverage infrastructure developed in the ELIXIR Tools Platform (https://elixir-europe.org/platforms/tools) and the EOSC-Life Tool and Workflow Collaboratory (https://www.eosc-life.eu/tools-workflows). The project will employ community-developed schemas and services improving the discoverability of workflows (such as WorkflowHub; see https://workflowhub.eu).

Applying FDOs and FAIR semantics and mapping services to all digital objects produced by the project will provide a harmonised and interlinked interoperability fabric for the BioDT infrastructure. Work will exploit the latest developments using RO-Crate to deploy FDOs for workflows and models in virtual research environment platforms such as Galaxy, LifeWatch ERIC's Tesseract, and Jupyter notebooks. The FDOs facilitate integration with the Common European Data Spaces and the European Open Science Cloud, and will lower the barrier for developers to compare and exchange models of Digital Twins and for researchers to use them.

The BioDT platform is intended to be a clean and lightweight system connecting existing and newly developed shared services for DTs with the planned users of the platform through well-defined APIs. For improved compliance with FAIR principles, the project will ensure that users as well as the platform and shared services support a common authentication and authorisation infrastructure solution, such as services provided by ORCID.

4.2 Scientific Papers and Other Project Outputs

Further to other types of data (Sections 2 and 4.1), BioDT project activities will generate outputs including scientific papers, posters presented at conferences, public presentations and deliverables.

In compliance with the FAIR principles and to ensure long-term preservation of project outputs also after the official end of the project, the project will use the multi-disciplinary repository Zenodo (https://zenodo.org) hosted by CERN in Geneva, Switzerland for storage of conference posters, presentations from public BioDT workshops and webinars as well for public deliverables created as a part of the project.

Zenodo accepts files in various formats and the documents will be stored in the same cloud infrastructure as research data from CERN's Large Hadron Collider using CERN's repository software Invenio (https://inveniosoftware.org), thus ensuring the long-term access, preservation and reusability of the material. Scientific papers will be published in open access journals. Each document deposited to Zenodo or to an open access journal will be assigned a Digital Object Identifier (DOI) to make them easily findable, citable and trackable.

Records of posters presented during conferences and presentations from public BioDT workshops and webinars will be linked from the BioDT Zenodo community (https://zenodo.org/communities/biodt) to the official project website (https://biodt.eu/). The project website will also include links to published scientific papers. In addition to Zenodo, all public deliverables will be published in CORDIS once they are approved by the EC Project Officer.

More information concerning the above-mentioned project outputs can be found in the BioDT Deliverable 2.1 Plan for Dissemination & Exploitation including Communication Activities which covers in detail the project engagement and outreach, exploitation and communication strategies.

5. Allocation of Resources

The responsibility for managing data related to the project lies with the following project members:

Responsibility	Responsible Party
Managing copies of internal communication and documentation, also after the project lifetime	CSC – IT Centre for Science Ltd
Data supplied as part of scientific publications, including depositing key data sets (including scripts and code) that support the publication in open repositories	Corresponding publication authors
Facilitating access to conference posters, scientific papers and training materials, with copies added to open repositories / databases where OA conditions are met	Trust-IT Srl
Access to primary biodiversity data (data to be reused by the project; see Section 2)	Research infrastructures as the primary data facilitators
Access to aggregated biotic and abiotic data sets, where not supplied as part of scientific publications or hosted by biodiversity RIs (with access requested by third parties)	CSC – IT Center for Science Ltd
Ensuring architecture components of the DT platform are FAIR-compliant, where possible	CSC – IT Center for Science Ltd and IT4Innovations National Supercomputing Center at VSB – Technical University of Ostrava

Table 5 – Data management responsibilities

Agreements with regard to FAIR distribution of software should also be followed, with details given in the project Consortium Agreement / Grant Agreement.

Improving quality of data, workflows and models through FAIR principles is within the scope of the project activities, thus the cost of these activities is included in the estimated budget of the project. The estimated budget of the project also includes funds for publication fees in full open access venues for peer-reviewed scientific publications. The availability of funds for FAIR research output management will be monitored by the Coordinator throughout the project.

The alignment of technical platform solutions for hosting and operating BioDT digital twins with those implemented by the DestinE initiative is intended to facilitate longevity of the project outcomes beyond the project funding period.

A sustainability model for the long-term preservation of BioDT will be developed as a part of the project, drawing from requirements from the research infrastructures, use cases and engagement with key strategic initiatives. The sustainability model will leverage existing arrangements at the European level, especially related to the evolution of EuroHPC and DestinE, and the insights gained from relevant project activities, notably from the integration with RI environments and collaboration with other initiatives and programmes. The sustainability model will assemble the business case for BioDT after the project official end covering, among other topics, the following: potential funding sources, market assessment and business models, operational and maintenance costs, and dealing in detail with relevant IPR aspects.

6. Data Security

6.1 Access to and Processing of Sensitive or Restricted Data

Where BioDT Use Cases require access to and processing of sensitive and/or restricted data (Section 3.2.2), handling of such data will be conducted in accordance with requirements established by the biodiversity RIs providing access to the data, and the providers of the original (raw) data. Sensitive and/or restricted data to be handled on LUMI will be pre-processed to ensure such requirements are met. Additionally, the LUMI supercomputer will feature solutions for sensitive data transfer and processing (e.g. based on data encryption, virtualisation and containers, as well as system-level procedures for job and file system isolation). Solutions for sensitive data processing on LUMI are prepared by the LUMI consortium, in collaboration with CSC – IT Center for Science Ltd and the ELIXIR community.

6.2 GDPR Compliance

As part of the BioDT digital twin platform being deployed on the LUMI supercomputer, the project will leverage an API-based access layer provided by the Puhuri project coordinated by the Nordic e-Infrastructure Collaboration (NeIC; https://neic.no). This access layer allows GDPR-compliant access to supercomputing facilities for pan-European collaborations, national research infrastructures as well as commercial entities.

In principle, re-purposed open biodiversity data do not contain sensitive personal information. In cases where the project requests detailed data sets from the original source or collects data from citizens as part of a citizen science project(s) or applications, transfer of personal data can only happen with the consent of the data originator, or the data must be fully anonymised. At all times, the project organisation ensures that the relevant data protection requirements are complied with and the statutory time limits for storage are met, and that all necessary technical and organisational measures are taken. Where the BioDT Use Cases involve the processing of GDPR-sensitive data, details on steps taken to ensure these requirements are met will be described and maintained in future versions of the BioDT DMP.

Details on data recovery and the secure storage, archiving and transfer of sensitive data are to be provided in a subsequent iteration of the BioDT DMP, as determined by the further development of BioDT Use Cases and infrastructure solutions supporting the operation of prototype DTs.

7. Ethics

An Ethics Appraisal Report completed on 12 January 2022 for the BioDT project identified no ethical issues of concern.

Research conducted under the BioDT project will comply with the highest ethical principles and standards of research integrity, including compliance with applicable EU, international and national laws.

The project partners will respect the fundamental principle of research integrity as set out in the European Code of Conduct for Research Integrity. This implies compliance with the following principles:

- Reliability in ensuring the quality of research reflected in the design, the methodology, the analysis and the use of resources
- Honesty in developing, undertaking, reviewing, reporting and communicating research in a transparent, fair and unbiased way
- Respect for colleagues, research participants, society, ecosystems, cultural heritage and the environment
- Accountability for the research from idea to publication, for its management and organisation, for training, supervision and mentoring, and for its wider impacts

The BioDT project partners ensure that persons carrying out research tasks follow good research practices including ensuring, where possible, openness, reproducibility and traceability, and refrain from research integrity violations described in the Code.

8. Intellectual Property Policy

The purpose of this Intellectual Property Policy (IPP) is to clarify general working principles with reference to the intellectual property aspects of the BioDT project, and to set recommendations regarding open source licensing. This policy is not meant to create new binding rules or principles (these can be found in the BioDT Consortium Agreement and Grant Agreement, which are visible and accessible only to the consortium members).

<u>Relation to the BioDT Consortium Agreement</u>

As per the BioDT Consortium Agreement, the project parties aim at making all results shared and placed under appropriate open licenses. Such licenses should be as open as possible. However, in case of software, the license must also be compatible with the requirements of the licenses of the software's dependencies so that the software can be legally combined. When results are published and made available under an open license, such a license shall be then applied to the exploitation of such results.

Exceptions to the general principles outlined in this IPP (such as deviations from the recommended software licenses) may be granted under provisions separately described in the BioDT Consortium Agreement. Where such separate provisions in the Consortium Agreement are made, they are to prevail over those described in this IPP.

<u>Principles and recommendations with reference to IP aspects</u>

The following principles and recommendations concerning IP aspects and licensing are set forth:

- A party of the BioDT project owns the IP created by it.
- The BioDT consortium will (non-exclusively) manage and exploit the IP for the benefit of the project.
- Where permitted by the license terms and conditions applied, software IP released as part of the BioDT project is to be licensed under an open source license.
    - Permissive licenses are recommended, but copyleft licenses can also be used, then it must be assured that it will not cause compatibility issues with other license terms.
    - Particular license recommendations include a) for software the i) MIT License or ii) Apache 2.0 License and b) for non-software the i) Creative Commons BY 4.0 license, ii) Open Data Commons Attribution License v. 1.0, iii) Open Data Commons Open Database License (ODbl) v.1.0, or iv) CC0 Public Domain dedication.

Principles and clarifications of Access rights

Results and Background licensed under an open source license do not need to concern themselves regarding the access rights, because these license terms fulfill grants of access right.

If not licensed under an open source license, it is recommended that parties make sure Access rights are granted according to the BioDT Consortium Agreement (Chapter 9). Full rules regarding Access rights are documented in the BioDT Consortium Agreement and Grant Agreement.

9. Other Issues

9.1 Internal Documentation of Project Materials and Files

The Coordinator of the BioDT project, CSC – IT Centre for Science Ltd, has created for the project partners a space for internal collaboration and information sharing. This space is located on Eduuni-wiki which is designed for creating content and sharing information in one place with the aim of making teamwork more organised and effective.

Eduuni-wiki is a browser-based service that is provided by CSC – IT Center for Science Ltd, and is used and maintained on the Internet. It is divided into separate sites that have their own URL and permissions. The BioDT project space is visible and accessible only to the persons working in the BioDT project.

Since it is not possible to edit Microsoft Office (Word, Excel, PowerPoint) documents directly on Eduuni-wiki, the Coordinator has made available a SharePoint library which is linked to Eduuni-wiki. This makes all needed information easily findable in one location.

For the purposes of the BioDT project, Eduuni-wiki is used to:

- make information about the project and its progress easily available for all in the BioDT consortium
- write and keep the agendas and meeting minutes of the Consortium bodies (Council, Project Management Board, Project Management Office and External Advisory Board) as well as those of the work packages
- keep track of action points, timelines and deadlines

For the purposes of the BioDT project, the SharePoint library is used to:

- prepare project deliverables, milestones and reports that require collaborative writing
- prepare joint presentations and publications
- store final versions of deliverables, milestones and other reports

CSC – IT Centre for Science Ltd as the Coordinator manages the access rights to the internal project collaboration space. No materials or files are foreseen to be deleted after the project official end. However, as Eduuni-wiki is not an official electronic archive, the registration and final storage of internal project records will be done according to the policies of CSC – IT Centre for Science Ltd.

10. Future Additions

Additions to subsequent versions of this DMP will include:

- Anticipated storage requirements (sizes) of data sets to be used and generated
- Further information on data sources, types and formats
- Requirements and steps for the processing and analysis of sensitive and/or restricted data
- Technical solutions for sensitive data storage, archiving and transfer, and data aggregation
- Further details on the BioDT sustainability model

Information on these topics will be added in future versions of the DMP in accordance with the development of BioDT Use Cases, digital twins and computing infrastructure solutions used to operate the BioDT digital twin platform.

Reference List

1. Wieczorek J, Bloom D, Guralnic R, Blum S, Döring M *et al.* 2012. Darwin Core: an evolving community-developed biodiversity data standard. *PLoS ONE* 7(1): e29715. DOI:10.1371/journal.pone.0029715
2. Jones MB, O'Brien M, Mecum B, Boettiger C, Schildhauer M *et al.* 2019. Ecological Metadata Language version 2.2.0. *KNB Data Repository*. DOI:10.5063/F11834T2
3. GBIF Secretariat. 2021. *GBIF Backbone Taxonomy*. Copenhagen: GBIF Secretariat. DOI:10.15468/39omei
4. *EnvThes - Thesaurus for long term ecological research, monitoring and experiments.* http://vocabs.lter-europe.net/EnvThes (Accessed 27 October 2022)