Developing Prototype Digital Twins for biodiversity conservation and management: achievements, challenges and perspectives

Damien Lecarpentier[‡], Timea Biro[‡], Dag Endresen[§,§], Marina Golivets[|], Volker Grimm[|], Sharif Islam[¶,#], Hanna Koivula[‡], Dirk Pleiter[¤], Tuomas Rossi[‡], Dmitry Schigel[«], Christoph Wohner[»], Gabriela Zuquim[‡], Jesse Harrison[‡]

‡ CSC - IT Center for Science, Espoo, Finland
§ University of Oslo, Oslo, Norway
| Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany
¶ Naturalis Biodiversity Center, Leiden, Netherlands
# DiSSCo, Leiden, Netherlands
¤ KTH Royal Institute of Technology, Stockholm, Sweden
« Global Biodiversity Information Facility - Secretariat, Copenhagen Ø, Denmark
» Environment Agency Austria, Vienna, Austria

Corresponding author: Gabriela Zuquim (gabriela.zuquim@csc.fi)

Academic editor: Editorial Secretary

Abstract

The BioDT project, funded by the European Union Horizon Europe Programme, seeks to investigate and push the boundaries of methodological, intellectual, data, technical and relevance aspects of the digital twin (DT) approach for understanding and predicting biodiversity patterns and processes. In this special issue, we present the reasoning and goals of the project and the overall advancements made towards the development of ten prototype Digital Twins (pDTs). Based on models of ecosystems, species and biological processes, digital twins integrate diverse data sources with advanced modelling and simulation techniques to provide a comprehensive, dynamic and data-driven understanding of biodiversity. Leveraging the EuroHPC LUMI infrastructure and data and expertise from GBIF, eLTER, DiSSCo and LifeWatch ERIC research infrastructures, the project aims to develop hybrid models combining the strengths of statistical and process-based models for improved predictive capabilities. While there are specific challenges and limitations to biodiversity DTs, the pDTs of BioDT lay the foundation for model-data fusions that will offer potential benefits to a wide range of end-users, from researchers and conservation organisations to policy-makers, land managers, educators and private sector companies. By fostering global interoperability in biodiversity data and research, BioDT paves the way for future collaborative efforts in biodiversity conservation and management.

Keywords

biodiversity, Digital Twin, modelling, research infrastructure, FAIR principles, High-performance computing

Introduction

A digital twin is a virtual representation of a physical object, system or process that simulates and is closely linked to its real-life counterpart. What distinguishes digital twins from mere models is their continuous updating with data for re-calibration and improving the model´s prediction. In addition, digital twins are interactive and provide software components that simplify model execution and interpretation for non-experts. In the context of biodiversity, a digital twin refers to a model that represents ecosystems, species or biological processes. It combines various data sources, including remote sensing, field observations, scientific collections and citizen science, with advanced modelling and simulation techniques.

While digital twins have been successfully implemented in engineered systems such as rockets or turbines, where all components and processes are well defined and understood due to human design, the complexity and limited understanding of natural systems pose significant challenges to the development of digital twins for biodiversity. Despite these obstacles, adopting the digital twin approach for biodiversity holds significant potential (de Koning et al. 2023).

This paper summarises the objectives, approach and results of the BioDT project in which a multidisciplinary team of experts in biodiversity research, ecological modelling, high-performance computing, artificial intelligence and FAIR data has been collaborating to prototype a series of digital twins. The project leverages the EuroHPC LUMI infrastructure and biodiversity data, as well as expertise from GBIF, eLTER, DiSSCo, LifeWatch ERIC and citizen science. While we provide an overview and summary, details about the specific experience made with each pDT are to be found in the corresponding articles of this special issue.

Overcoming the Challenges and Limitations of Modelling Natural Systems

Modelling biodiversity is a daunting task due to the vast biological complexity and dynamic nature of ecosystems and its components, data limitations and inherent uncertainty in ecological systems (Johnston 2024). These factors make it difficult to develop a comprehensive, accurate and up-to-date model with high confidence in predicting biodiversity patterns and processes (de Koning et al. 2023). Biodiversity modelling, therefore, focuses on specific systems, species and processes which cover key aspects of biodiversity. Hence, BioDT developed a whole set of distinct digital twins. Each of these digital twins has been designed to address a specific use case with limited scope, interactions and data, making the task more manageable and feasible (Table 1 and section “Ten Stories on Prototyping Biodiversity Digital Twins” below).

Our goal was to turn phenomenological (statistical) and mechanistic (process-based) into prototype DTs by establishing automated links to dynamic data sources, by leveraging the power of High Performance Computing (HPC) and by focusing on clearly identified end-users and their questions. Existing biodiversity models are usually based on static datasets, which are quickly outdated and also do not capture the response of the real systems to changing conditions. By automated updates from dynamic data sources, the model’s state variable can be updated, the models be re-calibrated and the process representations be improved, if needed. Additionally, by focusing on the end-users of the models, it is ensured that the model outputs are directly relevant for management and policy development.

DTs are also usually run for a larger region than the original model. This upscaling and also the need for repeated re-calibration at these larger scales, can require considerable computational power. BioDT thus relies on the EuroHPC LUMI supercomputer for high-performance simulation and data processing which allows computationally heavy models to perform in reasonable running time. Developing ten pDTs implies developing workflows for importing updated data in unified formats, including data such as weather or land-use data that are relevant for several pDTs. This fosters interoperability in biodiversity data and research, reducing fragmentation in the biodiversity research landscape and paving the way for future collaborative effort.

Prototyping Biodiversity Digital Twins

The BioDT project seeks to investigate and push the boundaries of methodological, intellectual, data, technical and relevance aspects of the digital twin approach for understanding and predicting biodiversity patterns and processes. The project sets out to explore the feasibility, workflows, partnerships and other visible and invisible aspects of developing digital twins in an area where, to our knowledge, few twins have been built before, for example, an app that allows us to predict migration routes of cranes (*Grus grus* , https://sensingclues.org/craneradar).

As we push these frontiers and keep up with methodological and technological advancements, the planned three-year duration of the project will end and eventually bring our current efforts to a close. The past two years of intensive collaboration and integration have already led to insights and achievements that go beyond traditional biodiversity modelling and monitoring. The present collection of forum papers in this special issue (Golivets et al. 2024) provides a snapshot of the work's state at the 67% mark. We hope that sharing these findings in a forum format will stimulate discussions and encourage the initiation of new and continuation of ongoing digital twin projects.

It is important to keep in mind that the 10 prototype digital twin (pDT) papers in this special issue (Table 1) represent a snapshot of work-in-progress, revealing differences in the level of advancement amongst the prototypes at the time of writing. For instance, some pDTs enable users to incorporate their own data (e.g. Honey Bees and Grassland Biodiversity Dynamics pDTs), while others rely solely on data accessible through research infrastructure (e.g. Invasive Alien Species and Crop Wild Relatives pDTs). These differences can be attributed to factors such as model availability at the beginning of the project, data standardisation and FAIRness. Sociological aspects, like the composition of pDT teams and user audiences, have also likely influenced the prototyping processes.

With adequate infrastructure and support, computing power is not a significant obstacle for developing biodiversity digital twins. By employing parallel processing schemes for biodiversity modelling codes, their capacity has been expanded to accommodate larger-scale data. However, the bottleneck lies in the scaling limitations of the models themselves, in particular their development for CPUs (Central Processing Units) instead of accelerators, such as GPUs (Graphical Processing Units), rather than a lack of computational capacity. An example on how to overcome this shortfall is the development of Hierarchical Modelling of Species Communities (HMSC; Ovaskainen et al. (2017), Rahman et al. (2024)) adjustments to run in HPC environment. The Hmsc-HPC code substantially broadens the practical boundaries of fitting HMSC models to large ecological datasets (Rahman et al. 2024). As a GPU-accelerated code, it is compatible with modern supercomputer architectures. During the project's final year, some pDTs may aim to progress further by scaling up across geographic, temporal and taxonomic gradients.

The ten prototype digital twins developed during the project cover four biodiversity thematic groups (Table 1) chosen to form a gradient of data complexity - from georeferenced organism occurrences at a given moment in time to biological interspecific interactions - and ranging from fundamental to applied contexts with strong policy connections.

Contrary to initial expectations, the progress of the pDTs by the end of the second project year was not directly related to data complexity or the nature of the modelling context. Instead, it depended on the model's readiness at the beginning of the BioDT project, as each pDT had a different starting point. Some pDTs were based on well-established, published models, while others relied on partially developed models requiring further refinement before transitioning to a digital twin approach. At present, the prototypes primarily focus on the regions where the modelling data originated. However, there are plans to expand the pDTs to cover broader geographic scales in the future.

For pioneering and prototyping projects like these, identifying the target audiences, their needs and expectations can be challenging due to the lack of previous references. Initially, indirect evidence was used to consider anticipated audiences and their requirements. However, as the prototypes were exposed to stakeholder workshops, outreach events and demos, their focus quickly shifted towards meeting real-world needs, mostly by developing a Graphical User Interface that provided specific outputs and access to the most relevant parameters and settings.

The prototypes on *Species response to environmental change* focused on European grasslands and Nordic forests management, production and dynamics. This group of prototypes considers various management practices, such as mowing and fertilisation, while balancing them with biodiversity conservation, which is usually associated with less intensive management. Additionally, the prototypes account for environmental conditions, habitat degradation, climate change, extreme events and their impact on productivity, biodiversity, migration, adaptation, conservation, management and monitoring. Therefore, the prototypes include data on functional or taxonomic groups like grasses and trees, along with rapidly responding indicator biological groups, such as birds and the predictors or variables associated with abiotic, socioeconomic, occurrences, ecological and demographic data. The audiences have been diverse, including researchers, farmers, regulatory decision-makers, forest managers, conservation agencies and practitioners, citizen scientists and monitoring agencies. The cultural ecosystem services pDT uniquely aims at addressing recreation, tourism, intellectual development, spiritual enrichment, reflection and aesthetic experiences of people, differentiating the general public with citizen-science potential from influencers of evidence-based management.

Similarly to the above group, the *Genetically detected biodiversity* theme addresses land-use and land-management changes, population growth, climatic factors, including extreme events, pests, plant diseases and food security. By utilising environmental DNA (eDNA) metabarcoding data, this approach sheds light on cryptic or less-studied organisms, such as fungi, bacteria and micro-invertebrates, thereby contributing positively to the taxonomic balance in BioDT. Furthermore, the incorporation of phylogenetic diversity offers a less conventional method for measuring biodiversity, considering evolutionary history for conservation planning and biodiversity management. The target audiences for this theme include agrobiodiversity researchers and agencies, molecular ecologists, monitoring initiatives, commercial enterprises, conservation agencies and biodiversity management organisations.

In the *Dynamics and threats from and for species of policy concern* group, the Invasive Alien Species pDT focuses on biodiversity changes where gains, rather than losses, can result in ecosystem degradation and impact human well-being. The aim is to provide projections of potential invasion locations and extents across Europe for both research and applied audiences.

Lastly, the *Species interactions with each other and with humans* group of pDTs addresses African swine fever, a transmissible virus affecting wild and domestic swine populations, relying on habitat, host, infection and treatment data. Additionally, a species of insect livestock, honey bees and their stressors are modelled in the context of pollination changes and the viability and productivity of honey bee colonies in various landscapes and under different management and climate-change scenarios. End-users can either choose a specific site and use the pDT to assess the combined effect of forage availability, extreme weather and, in the future, pesticide applications, or evaluate the general suitability of honey bees for the entire Germany. The intended end-users include researchers, beekeepers and beekeeping institutes, pesticide producers and regulators, farmers and their associations and policy developers at the EU level.

Mechanistic and Statistical Models

In manufacturing and logistics, where digital twins have been developed and widely used for decades, models represent systems designed and built by humans with well-understood components and mechanisms. This allows for the creation of mechanistic models that simulate system development over time. However, statistical models, particularly AI-based ones, are gaining popularity.

Compared to engineered environments, socio-ecological systems have many unknown components and mechanisms that are not fully understood. As controlled experiments are usually not possible, mechanistic ecological models are developed, but remain uncertain, making absolute predictions impossible. Instead, these models are used to rank scenarios or gradually improve understanding. Statistical models, such as species distribution models, play a crucial role in ecology, with thousands of applications and ready-to-use software available. Both model types play an important role in biodiversity modelling, as they have complementary strengths. Ideally, they would be combined in hybrid models, but this will not be possible in BioDT. There will, however, be statistical surrogate models of mechanistic models, for example, the BEEHAVE honeybee model, which capture, for a wide range of inputs and regions, the essential mechanisms via regressions or machine-learning approaches.

BioDT reflects this diversity: only four out of ten pDTs are based on mechanistic models (grassland, forest biodiversity, wildlife diseases and honeybees), which took at least a decade to develop, test and apply with large data demands. Other pDTs rely on established correlative modelling approaches like species distribution models (e.g. Crop Wild Relatives or Invasive Alien Species). Some pDT models were developed from scratch. Although AI-based models are not yet included in BioDT, they will play an increasingly important role in future DT development and several BioDT pDTs.

Making Biodiversity Models Fit for Advanced Computing

Typically, computational models used in BioDT use cases were run on a laptop and utilised no or little parallel processing. In order to enable scalable and extensive simulations on supercomputers with such models, we have demonstrated two generic approaches for several pDTs: containerising models and upscaling through data parallelisation.

Containerisation is a software technology for encapsulating software applications and their required computing environments. Containerising models simplifies model deployment on supercomputers, as containers can be prepared externally and contain complex environments that may not be readily available in HPC systems. Containers are also easily transferable across different HPC systems and cloud platforms, providing a clean, versioned environment that can be shared with other users to promote reproducibility (Kurtzer et al. 2017). Containerised models are currently demonstrated for Forest Biodiversity Dynamics, Crop Wild Relatives, Honeybee and Cultural Ecosystem Services pDTs and the container recipes are published under the BioDT GitHub organisation (https://github.com/BioDT).

Upscaling models can be achieved in many cases efficiently by managing parallelisation outside the model code using a task scheduler such as HyperQueue. This approach has currently been demonstrated for Honeybee and Grassland Biodiversity Dynamics pDTs and, in general, it is suitable for use cases that require processing large amounts of independent input data, parameters and/or scenarios that can be processed in parallel. The advantadges of such are that it is modular as the task scheduler can be changed as needed without changes to the model code and that it does not require extensive rewriting of scientific models. Scientists can continue using their existing codes and preferred programming languages, which facilitates the long-term sustainability of HPC usage.

The described parallelisation approach is limited to upscaling with CPU resources only if the model code supports GPU execution. While significant improvements in model runtime can be achieved by using GPU acceleration, it requires significant re-implementation efforts. Within BioDT, work on enabling GPU acceleration has been done in collaboration with other projects for the HMSC framework (Ovaskainen et al. 2017, Rahman et al. 2024) that is used in multiple pDTs as well as the wider scientific community.

BioDT Technical Platform

The BioDT technical platform is designed to support a diverse spectrum of DTs, reflecting the variety of pDTs rising from the BioDT use cases. Although the specifics of each pDT differ significantly, they share common high-level requirements: processing input and output data, running computational models and engaging with end-users through a graphical user interface or API. The BioDT technical platform facilitates these shared functions.

As an infrastructure development project, BioDT aims to create, develop and maintain a technical platform that allows users to access and utilise various biodiversity pDTs, leveraging the advanced computing capabilities of the EuroHPC LUMI supercomputer and other high-performance computing sites. A detailed explanation of the BioDT Technical Platform architecture is available in a separate deliverable (Kallio et al. 2023) and summarised on the BioDT website (https://biodt.eu/). The pDTs developed in the project showcase different applications of the BioDT Technical Platform, each accessible via a user-friendly graphical interface.

The primary user interaction point for the BioDT platform is through BioDT web apps (https://app.biodt.eu), developed using the open-source R Shiny framework (https://github.com/rstudio/shiny). The BioDT technical platform is built on open-source components, including LEXIS (https://opencode.it4i.eu/lexis-platform), HEAppE (https://github.com/It4innovations/HEAppE) and Waldur (https://github.com/waldur).

In brief, the LEXIS platform is used to manage data flows and launch computational executions in HPC or cloud environments. HEAppE middleware serves as a HPC-as-a-Service solution, enabling executions in the EuroHPC LUMI supercomputer and other integrated HPC environments, such as the EuroHPC Karolina supercomputer. The required authentication and resource allocation are handled by the Puhuri service (https://puhuri.neic.no/), based on the Waldur platform.

To ensure the sustainability and reusability of the technical platform services deployed for BioDT, the code integrations and extensions developed within the project are published in the upstream versions of the corresponding open-source software components. Other codes developed within BioDT, such as model codes and scripts, are released as open source under the BioDT GitHub organisation (https://github.com/BioDT), whenever possible and not already published elsewhere.

Promoting FAIR Data Principles for Improved Data Sharing and Reuse

The importance of adhering to the FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles (Wilkinson et al. 2016) in constructing efficient and sustainable biodiversity digital twins has been part of our strategy from the inception. Over the past decade, several consortium members have been actively involved in the FAIR landscape within the Biodiversity field, applying lessons learned from Biodiversity Information Standards (TDWG) and contributing to various EU projects such as ENVRI FAIR, BiCIKL and WorldFAIR. Each project and research infrastructure has developed its own approach to implementing the FAIR principles, providing valuable insights into data management and interoperability.

The digital twinning paradigm introduces additional challenges related to workflow management, model integration and deployment, which supplement the existing data management lifecycle that each research infrastructure addresses from its respective domain perspective. The key challenge is to integrate these activities, data structures and management into a unified framework (both technical and social) that can be effectively used, maintained and sustained. The successful implementation of FAIR principles often relies on key components that are specific to the domain and use case. For instance, biodiversity occurrence data have distinct metadata requirements and data structures compared to historical climate data or other direct measurement data. Despite common elements in understanding biodiversity ecosystems, such as species distribution, vegetation coverage and temperature, achieving domain integration and interoperability remains challenging.

Additional challenges include managing persistent linking and identifiers, common metadata structure and workflow management. Addressing these components is crucial, but as a whole, they impact the FAIR implementation for digital twinning. The concept of distributed infrastructure and federation adds another layer of governance and implementation complexity, along with FAIR implementation. The technical foundation of any digital twin service or platform must address this multilateral framework, where each partner and digital component is treated as an independent, sovereign entity with full control over its internal affairs. This collaborative approach aligns with the vision of Green Deal Data Space (https://green-deal-dataspace.eu/) and DestinE.

To create a cohesive FAIR implementation framework, the following components are essential:

1. Standardised data structures for consistency and easy integration.
2. Comprehensive model and metadata documentation for reproducibility, understandability and reusability.
3. Documentation of software dependencies following FAIR4ResearchSoftware guidelines (Barker et al. 2022).
4. Workflow integration and packaging datasets, models and software into machine-readable structures (e.g. RO-Crate).
5. Infrastructure integration of HPC and cloud resources for scalability and robustness.
6. Continuous monitoring and logging for transparency and accountability.

FAIR implementation ensures findability, accessibility, reusability, interoperability and sustainability of digital twins. This requires collaboration with data providers, proper indexing, clear provenance, integration across pDTs and maintaining well-documented components on platforms like GitHub.

Implementing FAIR principles is, therefore, crucial for establishing the essential building blocks required for developing and maintaining biodiversity digital twins. Without these principles, digital twins would lack the necessary structural robustness to facilitate intricate biodiversity research and conservation initiatives. It is also important to recognise that successful FAIR implementation necessitates expert support for data stewardship, model management and software management. Therefore, funding schemes, institutional support structures and domain research infrastructures should consider these aspects during planning. Capacity and skills building are equally vital. To optimise output reuse and alignment, each RI and EU project must collaborate and coordinate effectively. In addition to infrastructural and organisational readiness for FAIR, we witness a transition of professional and cultural norms, including within the BioDT project. This means for data management to shift from controlled by project or/and organisation towards open repositories, with national and global RIs enabling long-term data preservation and access to stimulate data upcycling.

Beyond Prototypes: Making Digital Twins Operational

The BioDT project aims to introduce the digital twin concept in the field of biodiversity and showcase its potential through a set of prototypes. A significant part of this initiative aims at devising various sustainability steps to transition these prototypes into fully operational digital twins.

The first step involved identifying relevant stakeholders, which can be any organisation or community that plays a significant and direct role in relation to a specific digital twin. This step includes pinpointing relevant end-users, determining the value that the digital twin will offer them, understanding how they will engage with it and conducting an initial evaluation of the user-base size. In this context, an end-user refers to a person who interacts with any of the digital twin services without being a service provider within the given twin system.

The possible end-users of digital twins resulting from the BioDT prototypes are diverse and will be of interest to anyone interested in understanding and addressing various critical issues in biodiversity conservation, management and climate change adaptation directly connected to the specific use cases considered in the project (Table 1). Examples of specific end-user groups relevant to BioDT include:

- Researchers and scientists studying Biodiversity dynamics;
- Conservation organisations planning and implementing strategies;
- Policy-makers and government agencies informing decisions and tracking progress;
- Land managers optimising resource use;
- Farmers and beekeepers, educators and students teaching and learning about biodiversity;
- Private sector companies assessing and mitigating environmental impacts;
- Citizen scientists contributing to and learning from research.

More generally, effective biodiversity conservation requires a comprehensive approach that addresses land-use changes, promotes sustainable practices and utilises innovative tools such as digital twins. EU initiatives, such as the EU Biodiversity Strategy for 2030, which aims to put Europe's biodiversity on a path to recovery by 2030, provide a framework for achieving these goals and promoting a more sustainable future for all. This strategy includes commitments to protect a certain percentage of land and sea, restore degraded ecosystems and enhance biodiversity in agriculture, forestry and urban and rural areas. On a global scale, the Convention on Biological Diversity, a multilateral treaty targeting the preservation of biological diversity, the sustainable use of its components and the fair and equitable sharing of benefits arising from genetic resources, presents another relevant framework.

The second step involves identifying relevant cost categories and examining future operational costs. For the majority of the BioDT pDTs, an initial understanding of the necessary computing and storage resources has been established. However, the actual requirements will largely depend on whether the currently often regional scope of the models will be expanded to a national and transnational level. Once a better understanding has emerged, cost coverage strategies can be formulated. It is expected that the BioDT digital twins will rely primarily on public funding, which is consistent with the fact that their operations are mainly a non-profit endeavour.

Finally, BioDT aims to align technical and data standards with research infrastructure, other digital twin initiatives and key strategic initiatives, such as the European Open Science Cloud (EOSC) and Destination Earth (DestinE). This alignment promotes synergy, prevents duplication and supports sustainability. Moreover, BioDT seeks to increase its uptake and use in research workflows by making its developments available on other European platforms and infrastructure, particularly through alignment with EOSC. Additionally, BioDT aims to contribute to strategic efforts and the biodiversity component of broader Earth system research through alignment with DestinE. Integrating BioDT with ecosystems like EOSC and DestinE can significantly benefit digital twins, as it facilitates data access, federation and discovery. A specific aim is to enhance data, software and community practice interoperability amongst the involved research infrastructures and integrate BioDT into relevant platforms, such as the LifeWatch ERIC-hosted BioDT Community Platform. This is accomplished through co-design, end-user hackathons and tailored training events.

This vision can also help research infrastructures supporting BioDT to establish themselves as data providers for various thematic and generic digital twins.

Conclusions

The development of biodiversity digital twins is a visionary idea turned into practice and has given its first steps towards a highly capable approach in predicting the behaviour and responses of different aspects of biological phenomena. At the beginning of the project, there was a high level of uncertainty related to the anticipated readiness of the outcomes given the intrinsic challenges of dealing with complex systems and the high heterogeneity of data sources, FAIRness and modelling strategies. Nevertheless, the pDTs under development in BioDT demonstrate multiple feasible ways in which the digital twin concept can be applied to the biodiversity field.

Several key takeaways have emerged from our experience in prototyping biodiversity digital twins. Firstly, successful collaboration is vital and requires the engagement of a diverse range of experts, including IT infrastructure specialists, modellers and biodiversity professionals. Secondly, it is crucial to leverage mature models instead of developing them from scratch to avoid expending significant resources early on, which could compromise the project's momentum and timeline. Thirdly, identifying end-users from the outset and developing digital twins with their needs in mind or through collaborative efforts (e.g. co-design workshops and hackathons) is indispensable. Lastly, interoperability, RI-based data flows and FAIR principles must be incorporated by design to ensure wider adoption, ease of use and support the reuse of components. The implementation of FAIR principles requires expert support for data stewardship, model management and software management, which must be available early on. In essence, creating digital twins for biodiversity is as much a human endeavour as it is a technological pursuit and recognising and addressing both aspects are essential for successful progress.

Acknowledgements

This study has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement No 101057437 (BioDT project, https://doi.org/10.3030/101057437). Views and opinions expressed are 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. *We acknowledge the EuroHPC Joint Undertaking and CSC – IT Center for Science, Finland for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC – IT Center for Science and the LUMI consortium, through Development Access calls* and Rahti and Pouta cloud services from CSC (Finland) and Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Afsar B, Eyvindson K, Rossi T, Versluijs M, Ovaskainen O (2024) Prototype Biodiversity Digital Twin: Forest Biodiversity Dynamics. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e125086
- Barker M, Chue Hong N, Katz D, Lamprecht A, Martinez-Ortiz C, Psomopoulos F, Harrow J, Castro LJ, Gruenpeter M, Martinez PA, Honeyman T (2022) Introducing the FAIR Principles for research software. Scientific Data 9 (1). https://doi.org/10.1038/s41597-022-01710-x

- Chala D, Kusch E, Weiland C, Andrew C, Grieb J, Rossi T, Martinovic T, Endresen D (2024) Prototype biodiversity digital twin: crop wild relatives genetic resources for food security. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e125192
- de Koning K, Broekhuijsen J, Kühn I, Ovaskainen O, Taubert F, Endresen D, Schigel D, Grimm V (2023) Digital twins: dynamic model-data fusion for ecology. Trends in Ecology & Evolution 38 (10): 916-926. https://doi.org/10.1016/j.tree.2023.04.010
- Frøslev T, Boyd R, Schigel D (2024) Prototype Biodiversity Digital Twin: prioritisation of DNA metabarcoding sampling locations. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e124978
- Golivets M, Sharif I, Wohner C, Grimm V, Schigek D (Eds) (2024) Building Biodiversity Digital Twins. Research Ideas and Outcomes. https://doi.org/10.3897/rio.coll.240
- Groeneveld J, Martinovic T, Rossi T, Salamon O, Sara-aho K, Grimm V (2024) Prototype Biodiversity Digital Twin: honey bees in agricultural landscapes. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e125167
- Ingenloff K, Aziza SB, Weiland C, Nikolova N, Thulke HH, Lange M, Reichold A, Schigel D (2024) Prototype Biodiversity Digital Twin: Disease Outbreaks. Research Ideas and Outcomes 10: e125521. https://doi.org/10.3897/rio.10.e125521
- Kallio A, Vancraeyenest A, Lazovik E, Livenson I, Martinovič J (2023) BioDT Architecture and Implementation Plan. BioDT Project Deliverable D3.1 https://doi.org/10.5281/zenodo.10912141
- Khan T, El-Gabbas A, Golivets M, Souza A, Gordillo J, Kierans D, Kühn I (2024) Prototype Biodiversity Digital Twin: Invasive Alien Species. Research Ideas and Outcomes 10: e124579. https://doi.org/10.3897/rio.10.e124579
- Mikryukov V, Abarenkov K, Jeppesen T, Schigel D, Frøslev T (2024) Prototype Biodiversity Digital Twin: Phylogenetic Diversity. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e124988
- Ovaskainen O, Tikhonov G, Norberg A, Guillaume Blanchet F, Duan L, Dunson D, Roslin T, Abrego N (2017) How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters 20 (5): 561-576. https://doi.org/10.1111/ele.12757
- Ovaskainen O, Lauha P, Lopez Gordillo J, Nokelainen O, Rahman A, Souza A, Talaskivi J, Tikhonov G, Vancraeyenest A, Lehtiö A (2024) Prototype Biodiversity Digital Twin: Real-time bird monitoring with citizen-science data. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e125523
- Rahman AU, Tikhonov G, Oksanen J, Rossi T, Ovaskainen O (2024) Accelerating joint species distribution modeling with Hmsc-HPC: A 1000x faster GPU deployment. bioRxiv https://doi.org/10.1101/2024.02.13.580046
- Rolph S, Andrews C, Carbone D, Gordillo JL, Martinovič T, Oostervink N, Pleiter D, Sara-Aho K, Watkins J, Wohner C, Bolton W, Dick J (2024) Prototype Digital Twin: Recreation and Biodiversity cultural ecosystem services. Research Ideas and Outcomes 10: e125450. https://doi.org/10.3897/rio.10.e125450
- Taubert F, Rossi T, Wohner C, Venier S, Martinovič T, Khan T, Gordillo J, Banitz T (2024) Prototype Biodiversity Digital Twin: grassland biodiversity dynamics. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e124168
- Wilkinson M, Dumontier M, Aalbersberg I, et, al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Table 1.

Table 1. Prototype Biodiversity digital twins in 2024. Ten publications in the BioDT collection in RIO that capture the state of the work on the prototype digital twins two years into the project https://doi.org/10.3897/rio.coll.240. One of the purposes of this paper is to serve as a preface and a summary of the ten use cases from four groups: https://biodt.eu/use-cases.

Biodiversity Topic	Title	System or targeted group	Model type	Scale	Ref. and DOI
Species response to environmental change	Prototype Biodiversity Digital Twin: Grassland Biodiversity Dynamics	Managed grassland	Photosynthesis driven individual-based model (GRASSMIND)	German	Taubert et al. (2024) https://doi.org/10.3897/rio.10.e124168
	Prototype Biodiversity Digital Twin: Forest Biodiversity Dynamics	Boreal forests	Joint species distribution models and LANDIS II	Finland	Afsar et al. 2024 https://doi.org/10.3897/rio.10.e125086
	Prototype Biodiversity Digital Twin: Real-time Bird Monitoring with Citizen Science Data	Bird communties in boreal forests	Joint species distribution models	Finland	Ovaskainen et al. 2024 https://doi.org/10.3897/rio.10.e125523
	Prototype Digital Twin: Recreation and Biodiversity Cultural Ecosystem Services	Cultural systems, for example, National Parks	ESTIMAP models and species distribution model algorithms	Cairngorms National Park, Scotland	Rolph et al. 2024 https://doi.org/10.3897/rio.10.e125450
Genetically detected biodiversity	Prototype Biodiversity Digital Twin: Crop Wild Relatives Genetic Resources for Food Security	Habitats of the germplasm of interest	Species distribution model algorithms (e.g. GAM, MaxENT)	Global	Chala et al. 2024 https://doi.org/10.3897/rio.10.e125192
	Prototype Biodiversity Digital Twin: Prioritisation of DNA Metabarcoding Sampling Locations	DNA-detected species	Spatial models and sampling density statistics	Denmark	Frøslev et al. 2024 https://doi.org/10.3897/rio.10.e124978
	Prototype Biodiversity Digital Twin: Phylogenetic Diversity	Phylogenetic variation in the taxonomic group of interest	PhyloNext	Global	Mikryukov et al. 2024 https://doi.org/10.3897/rio.10.e124988

Dynamics and threats from and for species of policy concern	Prototype Biodiversity Digital Twin: Invasive Alien Species	Plant invasive alien species	Joint species distribution models	Europe	Khan et al. 2024 https://doi.org/10.3897/rio.10.e124579
Species interactions with each other and with humans	Prototype Biodiversity Digital Twin: Disease Outbreaks	Wildlife epidemics (wild boar, African swine fever)	Individual-based spatially explicit mechanistic model	Europe	Ingenloff et al. 2024 https://doi.org/10.3897/rio.10.e125521
	Prototype Biodiversity Digital Twin: Honey Bees in Agricultural Landscapes	Honeybee colonies in agricultural landscapes	Individual-based spatially explicit mechanistic model (BEEHAVE)	German	Groeneveld et al. 2024 https://doi.org/10.3897/rio.10.e125167