Data Standards and Interoperability Challenges for Biodiversity Digital Twin: A novel and transformative approach to biodiversity research and application

Sharif Islam[‡,§], Hanna Koivula[|], Dag Endresen[¶,#], Erik Kusch[¶], Dmitry Schigel[¤], Wouter Addink[‡,«]

‡ Naturalis Biodiversity Center, Leiden, Netherlands
§ DiSSCo, Leiden, Netherlands
| CSC-TIETEEN TIETOTEKNIIKAN KESKUS OY, Espoo, Finland
¶ University of Oslo Natural History Museum, Oslo, Norway
# GBIF Norway, Oslo, Norway
¤ Global Biodiversity Information Facility - Secretariat, Copenhagen, Denmark
« Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands

Corresponding author: Sharif Islam (sharif.islam@naturalis.nl)

Abstract

The Biodiversity Digital Twin (BioDT) project (2022-2025) aims to create prototypes that integrate various data sets, models, and expert domain knowledge enabling prediction capabilities and decision-making support for critical issues in biodiversity dynamics. While digital twin concepts have been applied in industries for continuous monitoring of physical phenomena, their application in biodiversity and environmental sciences presents novel challenges (Bauer et al. 2021, de Koning et al. 2023). In addition, successfully developing digital twins for biodiversity requires addressing interoperability challenges in data standards.

BioDT is developing prototype digital twins based on use cases that span various data complexities, from point occurrence data to bioacoustics, covering nationwide forest states to specific communities and individual species. The project relies on FAIR principles (Findable, Accessible, Interoperable, and Reusable) and FAIR enabling resources like standards and vocabularies (Schultes et al. 2020) to enable the exchange, sharing, and reuse of biodiversity information, fostering collaboration among participating research infrastructures (DiSSCo, eLTER, GBIF, and LifeWatch) and data providers. It also involves creating a harmonised abstraction layer using Persistent Identifiers (PID) and FAIR Digital Object (FDO) records, alongside semantic mapping and crosswalk techniques to provide machine-actionable metadata (Schultes and Wittenburg 2019, Schwardmann 2020). Governance and engagement with research infrastructure stakeholders play crucial roles in this regard, with a focus on aligning technical and data standards discussions.

In addition to data, models and workflows are key elements in BioDT. Models in the BioDT context are formal representations of problems or processes, implemented through equations, algorithms, or a combination of both, which can be executed by machine entities. The current twin prototypes are considering both statistical and mechanistic models, introducing significant variations in (1) data requirements, (2) modelling approaches and philosophy, and (3) model output. The BioDT consortium will develop guidelines and protocols for how to describe these models, what metadata to include, and how they will interact with the diverse datasets. While discussions on this topic exist within the broader context of biodiversity and ecological sciences (Jeltsch et al. 2013, Fer et al. 2020), the BioDT project is strongly committed to finding a solution within its scope.

In the twinning context, data and models need to be executed within a computing infrastructure and also need to adhere to FAIR principles. Software within BioDT includes a suite of tools that facilitate data acquisition, storage, processing, and analysis. While some of these tools already exist, the challenge lies in integrating them within the digital twinning framework. One approach to achieving integration is through workflow representation, encompassing standardised procedures and protocols that guide the acquisition, packaging, processing, and analysis of data. The project is exploring Research Object Crate (RO-Crate) implementation for this (Soiland-Reyes et al. 2022). Implementing workflows can ensure reproducibility, scalability, and transparency in research practices, enabling scientists to validate and replicate findings.

The BioDT project offers a novel and transformative approach to biodiversity research and application. By leveraging collaborative research infrastructures and adhering to data standards, BioDT aims to harness the power of data, software, supercomputers, models, and expertise to provide new insights. The foundation provided by the data standards, including those of Biodiversity Information Standards (TDWG), is crucial in realising the full potential of digital twins, facilitating the seamless integration of diverse data sources and combinations with models.

Keywords

data integration, advanced modelling, FAIR, FAIR Digital Objects

Presenting author

Sharif Islam

Presented at

TDWG 2023

Acknowledgements

This study has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101057437 (BioDT project,). 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.

Funding program

HORIZON.1.3 - Research infrastructures

Grant title

Biodiversity Digital Twin for Advanced Modelling, Simulation and Prediction Capabilities

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Bauer P, Dueben P, Hoefler T, Quintino T, Schulthess T, Wedi N (2021) The digital revolution of Earth-system science. Nature Computational Science 1 (2): 104-113. https://doi.org/10.1038/s43588-021-00023-0
- 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 https://doi.org/10.1016/j.tree.2023.04.010
- Fer I, Gardella A, Shiklomanov A, Campbell E, Cowdery E, De Kauwe M, Desai A, Duveneck M, Fisher J, Haynes K, Hoffman F, Johnston M, Kooper R, LeBauer D, Mantooth J, Parton W, Poulter B, Quaife T, Raiho A, Schaefer K, Serbin S, Simkins J, Wilcox K, Viskari T, Dietze M (2020) Beyond ecosystem modeling: A roadmap to community cyberinfrastructure for ecological data-model integration. Global Change Biology 27 (1): 13-26. https://doi.org/10.1111/gcb.15409
- Jeltsch F, Blaum N, Brose U, Chipperfield J, Clough Y, Farwig N, Geissler K, Graham C, Grimm V, Hickler T, Huth A, May F, Meyer K, Pagel J, Reineking B, Rillig M, Shea K, Schurr F, Schröder B, Tielbörger K, Weiss L, Wiegand K, Wiegand T, Wirth C, Zurell D (2013) How can we bring together empiricists and modellers in functional biodiversity research? Basic and Applied Ecology 14 (2): 93-101. https://doi.org/10.1016/j.baae.2013.01.001

- Schultes E, Wittenburg P (2019) FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. Communications in Computer and Information Science3-16. https://doi.org/10.1007/978-3-030-23584-0_1
- Schultes E, Magagna B, Hettne KM, Pergl R, Suchánek M, Kuhn T (2020) Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. Lecture Notes in Computer Science138-147. https://doi.org/10.1007/978-3-030-65847-2_13
- Schwardmann U (2020) Digital Objects – FAIR Digital Objects: Which Services Are Required? Data Science Journal 19 https://doi.org/10.5334/dsj-2020-015
- Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández J, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022) Packaging research artefacts with RO-Crate. Data Science 5 (2): 97-138. https://doi.org/10.3233/ds-210053