Mobilising Long-Term Natural Environment and Biodiversity Data and Exposing it for Federated, Semantic Queries

Hanna Koivula[‡], Christoph Wohner[§], Barbara Magagna[|], Paolo Tagliolato Acquaviva d'Aragona[¶], Alessandro Oggioni[¶]

‡ CSC - IT Center for Science, Espoo, Finland
§ Environment Agency Austria, Wienna, Austria
| Go-FAIR Foundation, Leiden, Netherlands
¶ CNR-IREA, Naples, Italy

Corresponding author: Hanna Koivula (hanna.koivula@csc.fi)

Abstract

Biodiversity and ecosystems cannot be studied without assessing the impacts of changing environmental conditions. Since the 1980s, the U.S. National Science Foundation's Long Term Ecological Research (LTER) Network has been a major force in the field of ecology to better understand ecosystems. In Europe, the LTER developments are led by the the Integrated European Long-Term Ecosystem, critical zone and socio-ecological system Research Infrastructure (eLTER RI), a currently project-based infrastructure initiative with the aim to facilitate high impact research and catalyse new insights about the compounded impacts of climate change, biodiversity loss, soil degradation, pollution, and unsustainable resource use on a range of European ecosystems and socio-ecological systems. The European LTER network, which forms the basis for the up-coming eLTER RI, is active in 26 countries and has 500 registered sites that provide legacy data e.g., historical time-series data about the environment (not only biodiversity). Its site information and dataset metadata with the measured variables are available to be searched at the Dynamic Ecological Information Management System - Site and dataset registry (DEIMS-SDR, Wohner et al. 2019). While DEIMS-SDR data models utilize parts of the Ecological Metadata Language (EML) schema 2.0.0, location information follows the European INSPIRE specification.

The future eLTER data is planned to consist of site-based, long-term time-series of ecological data. The eLTER projects have defined eLTER Standard Observations (SO), which will include the minimum set of variables as well as the associated method protocols that can characterise adequately the state and future trends of the Earth's systems. (Masó et al. 2020, Reyers et al. 2017).

The current eLTER network consists of sites that differ in terms of infrastructure maturity or environment type and may focus on one or several of the future SOs or they are not yet executing any holistic monitoring scheme. The main objective is to convert the eLTER site network into a distributed research infrastructure that incorporates a clearly outlined mandatory monitoring program. Essential to this effort are the suggested variables for eLTER SOs and the corresponding methods and protocols for relevant habitat types according to the European Nature Information System (EUNIS) in each domain. eLTER variables are described by using the eLTER thesaurus "EnvThes". These descriptions are currently enhanced by the use of the InteroperAble Descriptions of Observable Property Terminology (I-ADOPT, Magagna et al. 2022) framework to provide the necessary level of detail required for seamless data discovery and integration. Variables and their associated methods and protocols will be formalised to enable automatic site classifications, by building on existing observation representations such as the Extensible Observation Ontology (OBOE), Open Geospatial Consortium's Observation and Measurement, and the future eLTER Standard Observation ontology.

DEIMS-SDR will continue to be used as a core service with an RDF representation of its assets (sites, sensors, activities, people) currently being implemented. This action is synced with the Biodiversity Digital Twin (BioDT) project to ensure maximum findability, accessibility, interoperability and re-usability (FAIRness; Wilkinson et al. 2016) of data through FAIR Digital Objects (FDO). Other (digital) assets such as datasets, models and analytical workflows will be documented in the Digital Asset Register (DAR) alongside semantic mapping and crosswalk techniques, to provide machine-actionable metadata (Schultes and Wittenburg 2019, Schwardmann 2020).

The Biodiversity Digital Twin (BioDT) project is bringing together biodiversity and natural environment data from seven thematic use cases for modeling. BioDT prototypes rely on openly available data that comes from multiple heterogeneous sources using a multitude of standards and formats. In the pilot phase, merging data requires "hand picking" from selected sources, and automation of workflows would still require many additional steps. There are ongoing efforts in both the BioDT and eLTER projects to find best ways and practices to bring the raw data together by using suitable standards but also to harmonise the other environment variables by referring to vocabularies and possibly express the data as FDOs.

Currently both the EML schema and Darwin Core standard (Darwin Core Task Group 2009; with registered extensions) allow referring to external schemas and vocabularies, which give flexibility but may still prove to be too narrow for the multitude of data types and formats the natural environment data requires. We welcome discussion about how to create good practices for enriching and harmonising natural environment data and species occurrence data in a meaningful way. GBIF's new data model and enriching the raw data with semantic artefacts may prove to be the way to provide thematic data products that combine data from multiple sources.

Keywords

natural environment data, EML, FDO, RDF

Presenting author

Hanna Koivula

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Darwin Core Task Group (2009) Darwin Core. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/450
- Magagna B, Moncoiffé G, Devaraju A, Stoica M, Schindler S, Pamment A, et al. (2022) InteroperAble Descriptions of Observable Property Terminologies (I-ADOPT) WG Outputs and Recommendations. Research Data Alliance. https://doi.org/10.15497/RDA00071
- Masó J, Serral I, Domingo-Marimon C,, Zabala A,, et al. (2020) Earth observations for sustainable development goals monitoring based on essential variables and driver-pressure-state-impactresponse indicators. International Journal of Digital Earth 13 (2): 217-235.
- Reyers B, Stafford-Smith M, Erb KH, Scholes RJ, Selomane O, et al. (2017) Essential Variables help to focus Sustainable Development Goals monitoring. Current Opinion in Environmental Sustainability 26-27: 97-105. https://doi.org/10.1016/j.cosust.2017.05.003
- Schultes E, Wittenburg P (2019) FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. In: Manolopoulos Y, Stupnikov S (Eds) International Conference on Data Analytics and Management in Data Intensive Domains. Communications in Computer and Information Science, vol 1003 https://doi.org/10.1007/978-3-030-23584-0_1
- Schwardmann U (2020) Digital Objects – FAIR Digital Objects: Which Services Are Required? Data Science Journal 19 https://doi.org/10.5334/dsj-2020-015
- Wilkinson M, Dumontier M, Aalbersberg I, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 https://doi.org/10.1038/sdata.2016.18

- Wohner C, Peterseil J, Poursanidis D, Kliment T, Wilson M, Mirtl M, Chrysoulakis N, et al. (2019) DEIMS-SDR – A web portal to document research sites and their associated data. Ecological Informatics 51: 15-25. https://doi.org/10.1016/j.ecoinf.2019.01.005