


Forum

Ecological Data Should Not Be So Hard to Find and Reuse

Timothée Poisot ^{1,*,@}
 Anne Bruneau,^{1,2}
 Andrew Gonzalez,^{3,@}
 Dominique Gravel,⁴ and
 Pedro Peres-Neto⁵

Drawing upon the data deposited in publicly shared archives has the potential to transform the way we conduct ecological research. For this transformation to happen, we argue that data need to be more interoperable and easier to discover. One way to achieve these goals is to adopt domain-specific data representations.

Open Ecological Data Can Accelerate Research

We are witnessing rapid changes in the practices designed to increase and improve the archiving and sharing of ecological data. These are essential steps toward preventing further loss of data [1]. Ultimately this new paradigm should facilitate massive data access and thereby generate novel research through data synthesis [2,3]. Presently, authors can deposit different kinds of information associated with a study: raw (or transformed) data, metadata, and, increasingly, scripts to reproduce the results of data analysis, short text descriptions, and output files. These data packages are hosted in general-purpose services or repositories (e.g., Dryad, figshare, Open Science Framework), code sharing platforms (e.g., GitHub, Zenodo, Bitbucket), or more specialized repositories (e.g., TreeBASE, GenBank, Morphobank, Open Tree of Life), and accessed either by following the link given in the original publication or by relying on the search abilities of the hosting services or data portals

such as DataONE. This model is designed to allow archiving first, and second, repeatability: researchers can get access to the original data, evaluate how they were analyzed [4], and, as such, increase their confidence in the original findings (repeatability). It also allows others to process data in different ways to assess whether the original results are robust to alternative methodologies, or test new hypotheses.

Well-Structured Data Are Difficult to Produce and Archive

Well-structured data can also be more easily combined with other datasets to conduct data syntheses, allowing the repurposing of existing data for exploring new and original questions that would otherwise be difficult to address, a task which is greatly facilitated by novel data processing pipelines [5]. Well-structured data also allow data-processing code to be more reusable (thereby saving an immense amount of time) and are easier to track and semantically annotate (thus facilitating provenance tracking and the attribution of credit). Currently, however, ecologists are either not aware or encouraged enough to use open, programmatically searchable, structured, specialized repositories for ecological data, which, when they exist, greatly improve current issues with data archiving for reuse purposes. One key issue is that many domains in ecology lack well-established, appropriate, and specific standards, as we discuss here. As a consequence, although ecological data are now more commonly deposited, they are not necessarily in a format conducive to their reuse, because researchers will favor deposition of data in the form in which they were collected or analyzed, or are unaware that a different format would be more conducive to reuse. It therefore appears crucial to initiate a dialogue between data producers, data reusers, and data managers to design data standards that would allow

widespread reuse and structured archival. This would also facilitate the evolution of more robust ways in which new formats can easily incorporate data based on previous ones.

The lack of domain-specific standards leads to a number of challenges faced by synthesis efforts that seek to extract novel knowledge from existing data. Roche and colleagues [6] determined that in journals with a strong data-archiving mandate, about 60% of data packages were published in a way that would prevent a complete or even partial replication of the original study, hindering both archival and reproducibility objectives. The modest rate of replication in ecological research [7] can be in part explained by difficulties in accessing raw data, but also by the lack of incentive to publish confirmatory studies in ecology [8]. Here we make the point that current data structure formats are not sufficiently standardized to allow researchers to organize their data in a way that would allow replicability and for reuse to become a more straightforward task. The current system is not fulfilling its potential to affect change in data archival practices, which suggests it needs improvement. Specifically, although the reproducibility of single studies has somewhat improved, integration of data from multiple studies remains a challenge: issues of reproducibility notwithstanding, the availability of integrated structured data from multiple sources would allow rapid implementation of novel research and testing of hypotheses at large scales.

Unstructured Data Impede Synthesis Research and Slow Down Training

Synthesis [9], and transdisciplinary synthesis involving ecological data in particular [10,11], is required to identify large-scale trends in biodiversity and ecosystem status and solve 'wicked problems' related to designing sustainability

practices and policies despite uncertain anthropogenic change [12]. Neither of these goals are easily attainable in a system where data aggregation can require hundreds or thousands of person-hours. More importantly, none of the tasks related to data reuse and aggregation can be automated to produce living documents, which can be updated in near-real time without human intervention. These living documents are an essential component if we are to synthesize our knowledge of biodiversity trends and document how our uncertainty changes. Due to the lack of interoperable formats, we are often limited to case-specific data aggregation efforts, with little hope for automation.

Ironically, massive data aggregation efforts are themselves difficult to replicate: they require identifying, accessing, understanding, and collating data from multiple heterogeneous sources, which makes synthetic datasets time consuming to build. There appears to be little added value for researchers to reproduce (or engage in) such studies. Similarly, access to different data sources is useful for training purposes, and particularly for courses aiming at familiarizing students with data-driven practices involving data manipulation and aggregation. We consider that activities involving data management and conservation for future usage in research has the potential, among other aspects, to improve data structure and analyses, develop organizational skills, and create collaboration opportunities at the early stages of researchers' careers. But combining data that are dispersed across many unstructured sources offers little assistance in generating skills for students who perform them. They are simply too time consuming; a lot of time is needed to understand how to retrieve, format, and reconcile disparate datasets, and in our experience students spend more time 'working on' rather than 'thinking about' the data.

The Way Forward Should Involve Greater Standardization

One promising way forward is to encourage and facilitate the use of open, programmatically searchable, structured, specialized repositories for ecological data. The Global Biodiversity Information Facility, which gives aggregated access to datasets on species occurrences and now totals over a billion records worldwide) is a shining example of what can be achieved when a specific data type (occurrence records) is hosted within a central repository, with a well-known and internationally accepted standard format [13]. Species occurrences, despite seemingly being a low-hanging fruit, require complex data specifications (i.e., Darwin Core Archive, which provides a controlled vocabulary for biodiversity data exchange) with multiple fields [14,15]. Scaling up, species traits, population time series, ecological interactions, environmental factors, among others, often need to be integrated with species occurrences to improve the understanding of ecological process and mechanisms and to test theory. Differences in measures and standards for these diverse ecological data types are such that developing an ecological data standard is not straightforward [14].

This suggests that we may not need a single unifying standard for all of ecology (as exemplified by Darwin Core or other specialized ontologies), but rather domain-specific data representations informed by real research use-cases, which can be made interoperable. The challenges of standardized data formats lie in creating a way of representing the many possible ways in which ecological data can be integrated, and finding which way is amenable to research work while ensuring the integrity of the information it represents. This requires a discussion at the interface between technical constraints and the needs of researchers, who act as both producers and consumers of data. Although adopting a domain-specific

perspective multiplies the burden of developing formats, it also makes each data representation more self-consistent as it aims to solve specific problems. These formats can then be coded and mapped onto a more general specification for access to data and metadata representation (for which the Ecological Metadata Language and the Darwin Core would be natural candidates). In this system, general (but at times unwieldy) data standards serve as a 'point of entry' into more precise or specific data representations, each serving a specific research domain.

Ecologists must not undertake the task of data standardization efforts alone; instead, computer and information scientists as well as librarians need to be involved [2], as is already the case in initiatives such as iDigBio or DataONE. Involving researchers and practitioners with diverse and complementary skills will facilitate and improve not only the design of data formats, but the establishment of a data publication process. Finally, we note that the current data formats and databases do not appear to be sufficient to ensure that data are easily accessible. Inadvertently, journals may have unfortunately contributed to the fragmentation of data by not enforcing the use of standards. Using standards would also open the possibility of automated data validation upon deposition, which would lessen the burden on reviewers by removing the need to check that data have been properly formatted and documented. We believe, however, that journals can contribute greatly by leading the way and making open data useful. This can be done by requiring the deposition of data in appropriate databases with a clear and documented format, which is inspired by research practices, whenever they exist. Societies and funding sources (e.g., agencies, government, research institutions, stewardship associations) for their part should support research and

education aimed at establishing data integration initiatives and improve data literacy. Funders, in particular, should also support the development, upgrade, operation, and maintenance of these structured databases. Universities can support their scientists in this effort by maximizing already available human resources in libraries to assist at all stages in the data publication process [2].

Concluding Remarks

Reproducibility and data integration are two key and complementary goals that should be achieved in the short term by deploying more structured data repositories. It is important to note that this does not require a radical and immediate shift in practices; instead, the ‘salvaging’ of unstructured data can be accomplished as an ongoing task, allowing a transition period. These goals should be a win–win situation for researchers and for society that directly or indirectly fund a large portion of the data that ecologists produce. Reproducibility, analyses, and extending the life of data beyond their original purpose via data formats that allow integration are key objectives that will lead to new and innovative research and provide the public with reliable and lasting scientific information. This note is a plea to academics and

journals to increase their efforts towards these goals and encourage the improvement of data archiving as well as the sharing of structured ecological data. Because data are the common currency for collaboration among ecologists, we call for a discussion around development of data standards involving researchers who collect data and researchers who reuse data. This process will involve many people and require a significant cultural shift in data storage and curation. Although challenging, we are convinced that this transformation will lead to a net gain overall for ecology as a science, by strengthening our capacity to meet society’s need for science-based solutions to a growing set of environmental problems.

¹Université de Montréal, Département de Sciences Biologiques, Pavillon Marie-Victorin 90, Avenue Vincent-d’Indy Montréal, (Québec) H2V 2S9, Canada

²Institut de recherche en biologie végétale, 4101 Sherbrooke Est, Montréal, (QC), H1X 2B2, Canada

³McGill University, Department of Biology, 1205 Docteur Penfield, Montreal, (QC), H3A 1B1, Canada

⁴Université de Sherbrooke, Département de Biologie, 2500 Boulevard Université, Sherbrooke, (QC), J1K 2R1, Canada

⁵Concordia University, Department of Biology, 7141 Sherbrooke Street West, Montreal, (QC), H4B 1R6, Canada

*Correspondence:

timothee.poisot@umontreal.ca (T. Poisot).

©Twitter: [@tpoi](https://twitter.com/tpoi) (T. Poisot) and [@bio_diverse](https://twitter.com/bio_diverse) (A. Gonzalez).

<https://doi.org/10.1016/j.tree.2019.04.005>

© 2019 Published by Elsevier Ltd.

References

- Gonzalez, A. and Peres-Neto, P.R. (2015) Data curation: act to staunch loss of research data. *Nature* 520, 436
- Renaut, S. *et al.* (2018) Management, archiving, and sharing for biologists and the role of research institutions in the technology-oriented age. *BioScience* 68, 400–411
- Zimmerman, A.S. (2008) New knowledge from old data: the role of standards in the sharing and re-use of ecological data. *Sci. Technol. Hum. Values* 33, 631–652
- Kenall, A. *et al.* (2014) An open future for ecological and evolutionary data? *BMC Evol. Biol.* 14, 66
- Borregaard, M.K. and Hart, E.M. (2016) Towards a more reproducible ecology. *Ecography* 39, 349–353
- Roche, D.G. *et al.* (2015) Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol.* 13, e1002295
- Evans, S.R. (2016) Gauging the purported costs of public data archiving for long-term population studies. *PLoS Biol.* 14, e1002432
- Nakagawa, S. and Parker, T.H. (2015) Replicating research in ecology and evolution: feasibility, incentives, and the cost-benefit conundrum. *BMC Biol.* 13, 88
- Baron, J.S. *et al.* (2017) Synthesis centers as critical research infrastructure. *BioScience* 67, 750–759
- Lynch, A.J. *et al.* (2015) Transdisciplinary synthesis for ecosystem science, policy and management: the Australian experience. *Sci. Total Environ.* 534, 173–184
- Prescott, M.F. and Ninsalam, Y. (2016) The synthesis of environmental and socio-cultural information in the ecological design of urban riverine landscapes. *Sustain. Cities Soc.* 20, 222–236
- White, R.L. *et al.* (2015) The next generation of action ecology: novel approaches towards global ecological research. *Ecosphere* 6, 1–16
- GBIF (2016) *GBIF Science Review 2016*, Global Biodiversity Information Facility
- Wieczorek, J. *et al.* (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One* 7, e29715
- Guralnick, R. *et al.* (2018) Humboldt Core – toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. *Ecography* 41, 713–725