Challenges in Curating Interdisciplinary Data in the Biodiversity Research Community

Inna Kouper[‡], Kimberly J Cook[§]

‡ Indiana University Bloomington, Bloomington, IN, United States of America
§ University of Kentucky, Lexington, United States of America

Corresponding author: Inna Kouper (inkouper@indiana.edu)

Abstract

Panelists: James Macklin, Agriculture and Agri-Food Canada; Anne Thessen, University of Colorado Anschutz Medical Campus; Robbie Burger, University of Kentucky; Ben Norton, North Carolina Museum of Natural Sciences

Organizers: Kimberly Cook, University of Kentucky; Inna Kouper, Indiana University

As research incentives become increasingly focused on collaborative work, addressing the challenges of curating interdisciplinary data becomes a priority. A panel convened at the TDWG 2021 virtual conference on October 19 discussed these issues and provided the space where people with a variety of experience curating interdisciplinary biodiversity data shared their knowledge and expertise.

The panel started with a brief introduction to the challenges of interdisciplinary and highly collaborative research (IHCR), which the panel organizers have previously observed (Kouper et al. 2021). In addition to varying definitions that focus on crossing the disciplinary boundaries or synthesizing knowledge, IHCR is characterized by an increasing emphasis on computation, integration of heterogeneous data sources, and work with multiple stakeholders. As such, IHCR data does not fit with traditional lifecycle models as it requires more iterations, coordination, and shared language.

Narrowing the scope to biodiversity data, the panelists acknowledged that biodiversity is a truly interdisciplinary domain where researchers and practitioners bring their diverse expertise to take care of data. The domain has a variety of contributors, including data producers, users, and curators. While they share common goals, these contributors are often fragmented in separate projects that prioritize academic disciplines or public engagement. Lack of knowledge and awareness about contributors and their projects and expertise as well as a certain vulnerability in branching out into new areas, are among the factors that make it difficult to tear down silos. As James Macklin put it, "... you're crossing a boundary into a place you don't maybe know a lot about, and for some people, that's hard to do. Right? It takes a lot of listening and thinking."

Due to their complex and interactive nature, IHCR projects almost always have a higher overhead in terms of communication, coordination, and management. Panelists agreed that for such projects there needs to be a collaboration handbook that assigns roles and responsibilities and establishes rules for various aspects of collaboration, including authorship and handling disagreements. Successful IHCR projects create such handbooks at the beginning and revisit them regularly. Another useful strategy mentioned was to hold debriefing sessions that evaluate what went well and what didn't.

Strong leadership that takes IHCR complexities into account and builds a network of capable facilitators and "bridge-builders" or "translators" is a big factor that makes projects succeed. Recognizing and encouraging the role of facilitators from the onset of the project helps to develop productive relationships across disciplines and areas of expertise. It also enables everyone to focus on their strengths and build trust.

Data and metadata integration is one of the big challenges in biodiversity, although it is not unique to it. Biodiversity brings together many disciplines and each of them identifies its own problems and collects data to address them. Data silos stem from disciplinary silos, and it will take a different, more integrated, kind of cyberinfrastructure and modeling to bring these pieces together. Creating such infrastructures and standards around interdisciplinary data and metadata are serious needs, although they are not valued and rewarded enough compared to, say publishing academic papers.

Lack of standardization and infrastructure also stands in the way of improving the quality of data in biodiversity. To evaluate the quality of data and to trust its creators, data users need to know who gathered and processed the data and how. When the data is re-used within a collaborative project, there is an opportunity to ask questions and find out why, for example, someone had certain naming conventions or processing and analytical approaches. Long-term data such as species' life history traits, however, can be collected over long periods of time. Improving the quality of biodiversity data requires going beyond interpersonal communication and addressing the issues of metadata and standards more systematically.

Panelists also discussed the issue of openness in connection to biodiversity data. Openness contributes to the improved quality of data and an increased return on public investment in science and research. Panelists' positions diverged in the degree to which biodiversity data should be open and approaches to address competitiveness and sensitivity in research. On one hand, they acknowledged the need for some form of embargo on data sharing to allow data originators to benefit from their effort; on the other, they argued that lack of openness promotes silos and diminishes the quality of research and its reproducibility. Panelists briefly discussed the COVID pandemic data as an example of how lack of openness and silos can be detrimental to finding solutions:

> "COVID has given us the best example we have of how silos do damage to things that could have gone better. ... the data wasn't available, if it had been open or not even necessarily open but had anybody had any idea that it existed somewhere, that would have helped a lot. … We are learning those lessons, governments are

> changing the way they do business because of it. And so for us, I mean, our community, I think this has been one of the best things that could have happened to us in some ways, simply because it forced a change of mindset. And it has forced citizens to get engaged." [James Macklin]

The panelists, who brought a wide range of expertise to the discussion, including semantic and digitization technologies, agricultural data, evolutionary biology, and mineralogy among others, discussed projects they work on, which engaged the audience and stimulated a discussion among all participants about the role of end users in biodiversity data curation, non-traditional careers in biodiversity, and approaches to reviewing data similar to traditional research publications. Panelists and the audience also discussed the differences between "cleaning" and "annotating" data, making annotations part of the biodiversity record and data reviews. These productive discussions provide a foundation for further developments in the research and practice of curating biodiversity data and building strong interdisciplinary communities.

Keywords

interdisciplinarity, collaboration, data curation

Presenting author

Kimberly Cook and Inna Kouper

Presented at

TDWG 2021

Funding program

Institute of Museum and Library Services, award # LG-246400-OLS-20, "Extending data curation to interdisciplinary and highly collaborative research"

Conflicts of interest

References

- Kouper I, Cook KJ, Nikolov D (2021) Changes and continuities in curating interdisciplinary data. RDAP Association Summit URL: https://osf.io/nkw6b/