Title: Governance and the Social Epistemology of Data Integration for Global Biodiversity Loss
Authors: Beckett Sterner and Steve Elliott

Researchers are collecting massive amounts of biodiversity data to monitor increasing extinction rates and population declines, which are harming ecosystems, their services, and human well-being (Pecl et al. 2017; Urban et al. 2016). These data are invaluable for guiding conservation and risk management efforts but are distributed throughout the world in thousands of data repositories. How should scientists and policy makers organize the process of data integration to address global biodiversity loss? We need new conceptual tools to address the social epistemic challenges posed by global data system interoperability to support integration. Leonelli (2019), for example, highlights data governance, which she defines “as the strategies and tools employed to identify, manage, and disseminate data,” and how such governance requires new accounts of evaluating research work beyond traditional academic metrics.

In this paper, we analyze how global data integration is reshaping how scientists produce collective knowledge about biodiversity trends. First, we show how centralized versus decentralized governance strategies take on complementary roles at different scales of the data integration process based on the different rights of actors with respect to accessing, using, and altering biodiversity data. Second, we argue that proposals for achieving global interoperability of biodiversity data systems must look beyond global requirements for supporting evidence- based decision-making and must consider their implications for the local authority and autonomy of individuals and subfields over collecting, representing, and valuing their data. Technical proposals for data integration are also social interventions that reshape and depend on the practices by which researchers know things collectively.

We begin by characterizing competing governance strategies for achieving global interoperability of biodiversity data. We employ the knowledge commons framework, according to which open biodiversity data comprise a shared pool of resources. We focus on two general strategies for making these data interoperable: centralization and decentralization. Countries have spent tens of millions of dollars to create international data aggregators that pull data from smaller repositories so users can access comprehensive datasets via central web interfaces. However, these efforts rely on centralized computational workflows that regularly introduce distortions and errors into the original datasets while providing few mechanisms for users to fix them. In contrast, smaller community data portals focus on limited thematic collections, e.g. by taxonomic group or spatial region, and typically engage scientific experts, enthusiasts, and conservation practitioners in contributing and curating data in a distributed fashion. These data portals typically do not share coherent taxonomic hierarchies or ecological metadata classifications, and they can differ in what data quality standards they apply.

We conclude that neither strategy, taken in isolation, is practicable. Rather, to fruitfully govern the shared pool to enable data integration, researchers should consider the importance of scale in deciding which aspects of their social organizations and of their repositories require centralization and which require decentralization. We summarize this result in a taxonomy of strategies, and indicate how the strategies can be empirically evaluated.

Pecl, G.T., et al. 2017. *Science* doi.org/10.1126/science.aai9214.
Urban, M.C., et al. 2016. *Science* doi.org/10.1126/science.aad8466.
Leonelli, S. 2019. *Harvard Data Science Review* doi.org/10.1162/99608f92.17405bb6.