Data Integration without Unification

Authors:
Beckett Sterner, Steve Elliott, Edward Gilbert, and Niko Franz

Correspondence:
Beckett.Sterner@asu.edu

Draft Date:
11 August 2020

Abstract:

Life scientists generate big data by pooling many smaller datasets and by ensuring that those datasets combine to form a trustworthy body of information with a net increase in use value. Most proceed by constructing a maximally comprehensive dataset based on universal standards for representing the data's empirical content and fit for different uses. We argue that this approach rests on an regulative ideal to create unified datasets, but following this ideal isn't necessary: there are alternatives that enable the benefits of data pooling to be realized through infrastructure supporting lateral exchange and customization of data among multiple sources. We illustrate data integration without unification in the context of big data for biodiversity, which aims to address rapid biodiversity losses across the globe.

Keywords:
Biodiversity loss, bioinformatics, Darwin Core, data science, data portal, data synthesis

1. Introduction

In the life sciences, researchers frequently arrive at "big data" by integrating many smaller datasets from different sources in a shared repository (Leonelli 2016, Berman et al. 2016). To realize the full benefits of this project for one or more research problems, scientists have to combine their data in ways that not only gets individual datasets to travel from one situation to another (Leonelli 2016), but also ensures that these datasets *integrate* to form a trustworthy body of information with a net increase in value for potential users (commensurate to the integration effort). Each data source typically reflects idiosyncrasies of labelling, pre-processing, sampling, and so forth, all of which are specific or local to the context in which the data were collected and first stored. As a result, the job of integrating and further curating data and metadata from many sources is a major challenge for data-intensive science. How can researchers overcome this challenge, and how can such integrated data be structured to foster their diverse research aims?

Big biodiversity data initiatives since the 1990s have generally assumed that the solution must involve building a dataset that is comprehensive (or "global") and based on consensus standards for valid data records and metadata categories (Bisby 2000, Godfray 2002; Peterson et al. 2010; Turnhout and Boonman-Berson 2011; de Jong et al. 2015; Ruggiero et al. 2015; Devictor and Bensaude-Vincent 2016). These standards largely determine the scope of questions or problems for which the dataset can be used. We call such a dataset a *unified dataset*. To enable the construction of unified datasets specific to their fields, many research communities maintain central repositories that store and enable access to deposited data through web-based portals.[1] The desire for a unified dataset located at a single point of access form a package has been a powerful motivator for the development of new data infrastructure initiatives (CITE dataONE, GBIF, iDigBio, GenBank, OpenTree).

We argue that the unification approach often rests on the false assumption that integrated data *must* result in a unified dataset if it is to overcome the challenges of idiosyncrasies in datasets from different sources. Despite criticisms of the unified approach in philosophy of science (e.g. Leonelli 2016, Franz and Sterner 2018), no one has articulated a comparably general alternative. We address this gap by introducing the concept of data pooling, which is a widespread scientific practice that consists of several components: bringing many distributed datasets into one (but not *only* one) place, provisioning adequate infrastructure to manage and provide access to them, and governing the resulting resource for the benefit of a community of users and stakeholders beyond a single research project or lab.

While a unified dataset in a central portal *can* be an outcome of data pooling, alternatives are possible, extant, and useful. In particular, data pooling can yield customizable datasets based on infrastructures that enable researchers to laterally exchange data while maintaining different local metadata standards. Such a reticulated strategy for data pooling can actually deliver certain

[1] Hereafter we refer to a repository along with its set of data, web site, governing organization, and management team simply as a *portal*, following common scientific practice.

benefits over the unified approach when experts need to represent conflict and ambiguity inherent in individual sources and versions, or emerging through the process of aggregation, e.g. when they disagree about how to characterize data but nonetheless wish to share data across projects. The core of our argument is that desires for unified datasets operate collectively as a regulative ideal in data pooling, that this ideal isn't necessary, and that other approaches to pooling data can achieve researchers' ends without the ideal of unification.

We illustrate approaches to pooling data with and without the unification ideal using examples from contemporary biodiversity data science. Biodiversity 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). To conduct this research at international and global scales, there are efforts to make unified datasets available via central portals, such as the Global Biodiversity Information Facility (GBIF), at the time of writing with nearly 1.6 billion individual, standard-compliant observations of species. Scientists generally refer to these observations as "species occurrence data" following Darwin Core, which is a set of standards for writing taxonomic metadata for occurrence records (Wieczorek et al. 2012). There are also efforts to pursue what we call the data pooling approach in ways that yield dis-unified datasets in multiple and dispersed portals that nonetheless enable useful sharing of data across those portals. An illustration of this latter approach is the class of several dozen portals constructed with the Symbiota software platform (Gries et al. 2014). Below, we illustrate these two approaches and show some of their strengths and weaknesses. Insofar as the pooling approach exemplified with Symbiota exists and is viable, it reveals the limits of the underlying assumption of the unification approach.

Our argument extends prior philosophical work on data integration. While most work on the topic of integration in philosophy of science has focused on explanatory integration, e.g. through models or mechanisms, several authors have recently noted the importance of data integration for scientific knowledge production and its benefits or risks to society (O'Malley and Soyer 2012, Leonelli 2013). "Often conceived as a problem or at least a major challenge, data integration is the activity of making comparable different data types from a huge variety of potentially inconsistent sources" (O'Malley and Soyer 2012, 61). Data integration happens on multiple scales, e.g. for individual projects and as part of building community-wide research infrastructure, and also over time as these systems continue to evolve in their contents and design. Data integration also forms a crucial part of integrative research more broadly that incorporates models, experiments, and both exploratory and hypothesis-driven approaches (O'Malley and Soyer 2012).

We focus on epistemic issues arising at the level of whole data ecosystems based on how scientists create shared pooled data resources (e.g. Gesing et al. 2017 and references therein).[2] Scientists who create and manage data repositories, for example, make decisions about the

[2] We treat data pooling as a form of data integration to match current language in philosophy of science as our primary audience. Biodiversity researchers often use "aggregation" instead. Some scientists distinguish data integration and aggregation based on whether one is combining data of different types or the same types, but this is at best a rough heuristic when facing messy and complex data.

number and connectivity of these resources the community can sustain, with widespread impacts for other researchers. For example, when scientists decide how to standardize the production and labeling of datasets, they can raise epistemic controversies about the extent to which whole research communities (and relevant stakeholders) must converge on shared ontological beliefs, methodological standards, and aims (Leonelli 2016, 2019; Sterner et al. 2020). Similarly, expert curators remain essential to the production of valuable big data by contributing to, enriching, and harmonizing the information in pooled databases. As a discipline, data-centric science struggles to reward and facilitate their work (Leonelli 2016, Franz and Sterner 2018).

To develop our account of data pooling, we pursue the following strategy. In the next section we review Todd Grantham's (2004) account of practical unification, which specifically functions to characterize unification outside of contexts of explanation. We show how Grantham's framework can accommodate data integration via unified datasets but that it only partially captures data pooling without unification. Next, we present our account of data pooling, which includes accounts of tunability and multiscalability as two criteria for portal-based digital datasets. These criteria can be used to evaluate portals for their capacities to deliver pooled datasets that are valuable to different audiences of users. In section 4 we review the unification approach to data pooling in biodiversity data science, and we show how it yields metadata standards and central portals that struggle to satisfy those two criteria. And in section 5 we illustrate an approach based on integration without unification and show how it yields metadata standards and portals that satisfy the two criteria.

2. Conceptualizing data integration without unification

The goal of building a comprehensive or global dataset may seem closely allied with the ideal of unified science. Occurrence data in biodiversity science, for example, are used to test theories across a number of scientific fields, including the study of the taxonomy, biogeography, and phylogenetic relationships among species. Occurrence records typically combine physically or digitally vouchered records of species observations with metadata that provide additional information about that record. Hence the voucher can be a photograph or physical sample of an organism, while the accompanying metadata include information relevant to the observation, such as the geographical coordinates, date, taxonomic classifications, study methods, and collection to which the records belong. Occurrence data are also essential to informing applied decision-making based on models of extinction risk in conservation biology, ecosystem function and resilience in ecology, and ecosystem services in environmental economics.

However, a global dataset of species observations doesn't represent unification in the sense of the reduction of higher-level scientific theories to more fundamental, lower-level ones. Indeed, all of the fields we mentioned above recognize organisms as ontological units but use information about them to address problems at different compositional scales and time frames. We therefore focus in this paper on how unification functions as a regulative ideal when issues of

reduction are not at stake. Grantham's (2004) account of unification conceptualizes it as interconnectedness among scientific fields, so it's especially suitable in this regard.

Grantham's account aims to characterize how a range of scientific activities beyond intertheory reduction can advance unification. He distinguishes two classes of interfield unification, theoretical and practical. "Fields can be theoretically unified as the intra-field theories become more densely interconnected. Fields can be practically unified insofar as one field comes to rely on the methods, heuristics, or data of a neighboring field" (Grantham 2004, 143). Theoretical unification can increase through new explanatory, ontological, or conceptual connections such as explanatory reduction, part-whole or causal relationships, or conceptual refinement. Practical unification, by contrast, increases through establishing dependence among activities and resources across fields, such as through using theories or methods from one field to generate new hypotheses in another or methodological integration that uses data from multiple fields to test hypotheses. Grantham emphasizes that unification is a matter of degree based on the interconnectedness among fields along both theoretical and practical directions.

We note several ways that the goal of producing a global dataset *can* lead to unification on Grantham's account. In terms of theoretical unification, pooling datasets from many sources benefits from the use of standardized classificatory theories (Leonelli 2016). These theories express general claims about the existence and properties of many entities or processes in the world. They also express claims about relevant or necessary information scientists should provide in measuring or manipulating their objects of study. With occurrence data, for example, the Darwin Core standard mandates inclusion of a taxonomic name for a valid record, and the Humboldt Core provides optional categories for describing a field study's sampling effort, e.g. opportunistic collection versus systematic species inventory in an area (Wieczorek et al. 2012, Guralnick et al. 2018). If researchers reach a global consensus of views about what there is to observe and how to observe it, then on Grantham's (2004) account they have achieved a strong unification of the concepts, methods, and beliefs involved in making and using the data.

For practical unification, a global dataset provides a list of all relevant data records for researchers to access across different fields, and it links researchers' analyses to a common source. To the extent that scientists can add new information to existing records, the global dataset may also serve to collect the observational and analytical outputs of many fields within a single resource. But one aim and consequence of standardization is to exclude some possibilities in favor of enabling others. So both the practical and theoretical unifying effects of a global dataset have the consequences of prohibiting or at least obstructing some research aims (Bowker 2000). Constructing a global dataset therefore has further consequences for the unity of ontology, methods, and importantly the aims among the fields involved.

What, then, might data pooling without unification look like? While Grantham treats unification as a matter of degree, this move elides important differences between cases in which moderate levels of interconnectedness reflect transient states on the way to total unifications, and cases in which moderate interconnects reflect stable, desirable outcomes. In talking of integration without unification, then, we are interested in why unification isn't a universal ideal for scientific

practice (Brigandt 2010). Philosophers have given a number of strong reasons for the absence of that universality in the context of explanatory integration, such as the value of using incompatible representations to study complex phenomena (e.g. Mitchell 2003). Brigandt (2010) notes that the goal of integration is often localized to solving particular research problems rather than achieving a general ideal of unified science.

3. Data pooling as a form of integration

The concept of "data pooling" provides a useful tool to characterize how scientists collect datasets from multiple sources without prejudging the appropriate scale or unity for that effort. As we define it, data pooling is a combination of bringing many distributed datasets into one place, provisioning adequate infrastructure to manage and provide access to them, and governing the resulting resource for the benefit of a community of users and stakeholders beyond a single research project or lab. The problems created by big data projects are often as much social as technical in character, requiring major changes to the cultures, organizations, and infrastructures of the research communities involved (Bowker 2000, Baker and Millerand 2016, Leonelli 2016). We characterize data pooling as a distinctive form of data integration in terms of its outcomes for shared data infrastructure and governance. While researchers can pool data in centralized portals so as to create unified datasets, that approach isn't the only option available.

Most data integration activities still result in private datasets held by specific labs or collaborative projects with no or limited access for people beyond the current or future members of those groups. For example, many projects integrate datasets from different model species to study a causal mechanism in a single target system, but they don't then publish the combined dataset to an online repository. Even those datasets are submitted often exist as "flat files" with no annotations linking data fields to shared metadata categories.

A necessary feature of data pooling is the provision of infrastructure sufficient to preserve the resulting data collection as a reusable resource across the lifespans of multiple projects, or perhaps indefinitely. Actors may secure adequate infrastructure by adopting existing tools and services or by creating custom technologies and institutional arrangements.

Contemporary examples of socio-technical infrastructure for data pooling include but aren't limited to data portals. Data portals include a repository for storing data, the pooled data and metadata themselves, an online Web interface for querying and retrieving stored records, an organization for administering the portal, a set of people filling roles within the organization, and a set of formal and informal institutions to govern the portal, such as metadata standards, co-operative agreements with other organizations, and job descriptions. As sites for social interactions, portals constitute boundary objects (Star and Griesemer 1989) that organize data practices among participants situated in different social worlds (Gerson 1983). People involved in running or using a portal often do things besides store and curate data, such as training new users, hosting workshops, conducting and publishing research, and helping to credential new researchers. Portals also have material inputs, such as funding and physical infrastructure

including preserved materials. Pooling data into a shared resource thus involves more than simply producing a new unit of collected information and incorporates social and economic components found elsewhere in scientific practice.

To characterize the pros and cons of alternative approaches for pooling data, we start by introducing two general evaluative criteria. First, the multiscale criterion refers to maintaining high accuracy of data record labeling across fine to coarse-grained partitions within any given metadata category. The criterion requires that data are pooled in a way that preserves the value of subsets of the total resulting dataset. For example, consider a metadata category with hierarchical structure, such as a taxonomic classification. A set of data records satisfy the multiscale criterion for a metadata category if they are annotated with few false negatives and positives at all levels of the hierarchy (Guala 2016, Remsen 2016). Similarly, if location information is annotated in different ways across data records, it would not satisfy multiscale integration to only preserve the coarsest-grain level of spatial resolution shared across all the sources.

Second, the tunable criterion refers to whether it's possible to generate pooled datasets that are fit for use to answer a range of types of problems. Most occurrence data, for example, lack information about how systematically the data collectors searched a field site for different species. This information is crucial for determining whether a dataset tells us which species were absent from an area, or only which ones were present. Similarly, is the dataset appropriate for use to estimate relative abundance of species in an area, or only that some individuals exist there? The Humboldt Core metadata standard we mentioned earlier provides categories for annotating occurrence records with information about sampling effort to increase their discoverability and fitness-for-use in different modeling problems (Guralnick et al. 2018).

The value of multiscale, tunable data integration reflects the variability in the features of localized phenomena. Take for example biodiversity research, for which variability reflects what's happening to species and ecosystems locally as a result of continued economic development and climate change. A large majority of users want data for only a subset of taxa, e.g. one genus of plants or family of insects as recognized in a pragmatically accepted classificatory framework, and many focus on specific spatial regions rather than on the whole planet. Hence one way to frame what infrastructure for biodiversity data should be able to do is to deliver all and only the data relevant to the scale, taxonomic and spatial, desired by users. Moreover, relevance is not limited to the scope of a user's query: it also involves each data record's fitness-for-use. Expert taxonomic identification under a coherent classification, for example, is an important factor for the credibility of models predicting future species ranges (Araujo et al. 2019). Relevance should then also include considerations of fitness-for-use. In practice, efforts to pool biodiversity data into large, centralized databases have encountered problems with these criteria, which we further describe in the next section.

4. Biodiversity Data Pooling: Challenges to Unification

Current efforts to generate big biodiversity data exemplify the features of data pooling discussed above and motivate the importance of data pooling practices for broader philosophical inquiry. If scientists are to address rapid global biodiversity loss, they require major socio-technical innovations to support data pooling on multiple scales of spatial and temporal resolution (Hobern et al. 2019). Inspired by big science efforts like the Human Genome Project, many countries and international organizations now aim to combine all data about where and when different biological entities—typically species—are located (Guralnick et al. 2007). Global data pooling efforts have already transformed the biodiversity community by producing new data resources and infrastructures as well as long-term organizational bodies responsible for sustaining and governing them. Dozens of biodiversity data portals, for example, now exist with formal organizational and governance structures, and with missions aimed at benefiting basic science research, policy or conservation decision-making, and public access to scientific knowledge (Gadelha Jr. et al. 2018).

Historically, the dominant strategy for pooling biodiversity data has been to build comprehensive databases and centralized access services, typically in the form of centralized national or global web portals and unified metadata systems (Guralnick et al. 2007, Baker and Millerand 2016). These portals provide centralized access to datasets sourced from a wide range of theses repositories, collections, and databases. Take for example the Global Biodiversity Information Facility (GBIF), which provides a portal administered by an international organization funded by member nations. It began in 2001 upon a recommendation from the Biodiversity Informatics Subgroup of the Organization for Economic Cooperation and Development's (OECD) Megascience Forum. GBIF now serves more than 1.5 billion occurrence data points for users to search and download, aggregated from a wide range of citizen science projects, museum collections, and other organizations. The dominant source for such occurrence data have been natural history collections and ecological surveys hosted and conducted around the world.

Biodiversity researchers have identified multiple major problems generated by the centralization required to produce a global dataset. We discuss three such problems: disagreement between portal and taxon-specific experts that can stifle debate; poor quality of pooled data; and lack of capacity to improve data quality. Due to these problems, central portals that aim to create unified datasets struggle to pool data in such a way that the latter are tunable and scalable.

First, central repositories impose universal standards for data classification, even where the relevant classificatory theories are unstable over time or lack consensus. These standards ostensibly enable the data records to be sorted so they can be searched and retrieved according to user queries. In the case of taxonomy, for example, GBIF authors a “taxonomic backbone” that effectively competes with the hypotheses of experts in systematics while at the same time presenting its portal as the main point of access to all biodiversity data (Franz and Sterner 2018, Garnett et al. 2020). Furthermore, pooling data for all taxa across the Earth imposes serious computational complexities that limit GBIF’s ability to harmonize fine-grained relationships

between datasets annotated under alternative taxonomies. In addition, imposing universal standards can weaken the relevance of those datasets to local or regional situations (Han et al. 2017).

Next, after several decades of efforts to build a centralized and unified global infrastructure, widespread deficiencies in data quality remain (Mesibov 2013, 2018, Franz and Sterner 2018). One illustration of these deficiencies comes from a recent publication on global species distributions of plants using the Botanical Information and Ecology Network (BIEN) database (Enquist et al. 2019). Out of an initial 200 million records in the database, the study discarded 165 million (83%) due to data quality problems with geocoordinates, taxonomic classification, and needing to exclude records about cultivated plants.

Third, pooling data in centralized portals doesn't necessarily help these data quality challenges, because the resulting databases typically lack the needed capacities. Centralized biodiversity portals typically restrict curation privileges for users and either outsource editing of records to other sources (e.g. GBIF) or allow participants to edit a subset of data (e.g. one museum collection or only the data people have contributed themselves). These restrictions impact the accuracy of data across scales and its tunability for different applications. Restrictions on editing centralized datasets can arise for a variety of reasons, including constraints from original data sources on modifying their content or the difficulty of vetting expert users on that scale. Users with corrections or new annotations must then contact each original data source individually in order to request edits. In addition, most biodiversity experts simply do not work on all groups and at the global level. Instead, the experts and communities tend to have both taxonomic and geographic (or even political) boundaries more accurately represented at low or middle level scales. Conversely, research communities that *are* interested in analyzing all organisms at the global level tend to lack the expertise needed, for example, to reconcile non-congruent classification schemes inherent in biodiversity data packages aggregated from multiple localized sources and communities of practice. The result is a primarily one-way flow of new and edited data from many distributed sources toward the global dataset with little gain in the community's net ability to collaborate on data curation and preserve improvements across many individual projects.

5. Biodiversity Data Pooling: Integration without Unification

Here we describe an alternative approach to pooling biodiversity data that does not aim to produce a single, global database. As a case study, we examine the ecosystem of biodiversity data portals created using the Symbiota software platform, and we assess its current capacity and future potential to deliver the benefits of multiscale, tunable biodiversity data. In geographic and taxonomic scale, the current Symbiota portals are regional rather than globally, but together they provide an example of a comprehensive strategy to provide the full range of multiscale, tunable datasets we've described. (Also see Campbell et al. 2020 for an example based on regional citizen science survey data.)

The Symbiota platform, as a technological artifact, is open-source software package that includes code for: a database schema, a content management system for user roles and metadata and a default human/computer interface, and a series of modules for ingesting and annotating data records in the repository (Gries et al. 2014). This software, designed and maintained by a team primarily at Arizona State University starting in 2008, enables users to create portals that provide online access to a biodiversity database and manage them in a decentralized fashion using a shared Web interface. Symbiota software also enables users to query the database via text or map searches and to create customized species checklists, often used by biodiversity researchers and enthusiasts. Researchers have launched over 30 biodiversity portals using the Symbiota platform, and they collectively host over 55 million records and receive tens of thousands of unique web visitors each month.

We highlight a few key features of Symbiota that enable lateral sharing of data among portals while tracking provenance and editing rights. The software currently allows two states for datasets pooled within a portal: (1) "live-managed," which means that the entity owning the physical collection of specimens or vouchers has comprehensive rights within the portal to create new occurrence records and annotations; and (2) "snapshots," which can be time-stamped versions of a live-managed collection exported to one or more outside portals where occurrence records may be annotated further – typically by actors who are not members of the entity that owns the physical collection.[3] The distinction between live-managed and snapshot datasets is essentially a matter of data governance, i.e. tracking where editing rights are vested within versus outside a portal.

In general, an individual data collection typically only undergo live-management in one (internal) portal, but may be represented as a partial or full snapshot in one or more external portals. Snapshot collections can be periodically (in some cases automatically) updated from respective the live-managed portal. Conversely, annotations made on snapshot occurrence records can be integrated with the corresponding live-managed collection under proper social and technical conditions. Accordingly, SEINet, a herbarium-based portal focused on North American vascular plants, can reciprocally exchange occurrence records and annotations with the National Ecological Observatory Network (NEON) portal, which includes taxonomically heterogeneous data sampled at the North American continental scale and is intended to facilitate long-term monitoring and forecasting of ecological macrosystems. Data managers for any Symbiota portal can therefore pool selected data from multiple sources into the same portal without losing differential control over the curation process.

How do these features of Symbiota make multiscale, tunable data pooling possible without unification under a global dataset? First, each portal makes its data available through a common standard for publishing to the Web (i.e. making a file accessible in a standardized

[3] Currently about 10-20% of snapshots are data imports from another live-managed Symbiota dataset, but this type of Symbiota-Symbiota data mapping is increasing. The majority of snapshot datasets are imports from an institution's local collections management software, such as Specify or Emu.

format on a public URL). Any other portal can then construct a customized dataset by importing partial or full snapshots from any source, not limited to Symbiota portals, that makes data available online in a Darwin Core compliant format. This importing process facilitates custom mapping of records where datasets differ in their metadata categories, or portal data managers can locally edit snapshot data (though these edits currently have to be reapplied after future updates, which can be an automated process).

Portals that use the Symbiota platform have the capacity to overcome the three problems associated with global portals described in the previous section. First, each portal maintains autonomy to select and curate pooled data according to metadata information *tuned to its aims and classificatory theories*. No universal taxonomic classification or other classificatory theory is necessary, as data harmonization only needs to happen between a targeted, often small number of sources. Next, Symbiota portals have the capacity to curate data locally. Compared to global portals, larger percentages of Symbiota portal users can contribute to the governance of metadata standards and store versions of cleaned data. As a result, and third, it's possible for portal users to maintain accurate data at the level of individual species and biodiversity monitoring projects.

For this approach to data pooling to work, at least two main conditions must be met. First, the producers of classificatory theories for occurrence data must provide sufficient information to enable translation across alternative theories (Sterner et al. 2020). The flexibility to customize pooled data across competing or historical viewpoints depends on the capacity to accurately map data records annotated under one system to the categories posited by another system. In principle this is consistent with constructing a universal classificatory theory so long as the ability to crosswalk data to other (perhaps less comprehensive) theories is maintained, but not by harmonizing data to a universal view that effectively replaces other theories. Second, data pooling projects work when they are closely aligned with communities (professional and enthusiast) with the expertise and resources to curate the resulting datasets. The point is not that everyone should get their own arbitrarily idiosyncratic dataset; rather, constructing a single, comprehensive dataset does not erase substantial differences in viewpoints and practices among fields. From an institutional perspective, enabling each community to take leadership and ownership of its data is a powerful incentive that is lost when control transfers to a single, global repository.

6. Conclusion

The rhetoric of the big data movement makes the project of scaling up datasets central to scientific progress. Sometimes scaling up is associated with reaching a new “natural” level, e.g. comprehensive information about species across the whole planet. Achieving such a global scale of data pooling, however, doesn’t eliminate demand for datasets scaled and attuned to local problems that scientists and decision-makers need to address. Regardless, the regulative ideal to build global unified datasets has operated widely throughout biodiversity data science. We reviewed how adhering to that ideal generates several significant problems. These include

fostering disagreement between portals and taxon-specific experts that can stifle debate; poor quality of pooled data; and lack of portal capacity to improve data quality. We suggest that the ideal for global unified datasets built on consensus metadata standards isn't a requirement for data science in biodiversity or other life sciences.

As a positive alternative, we propose a more general regulative ideal for pooled datasets. Such datasets result from several practices. First is bringing many distributed datasets into one place. This place can be a portal, but importantly there need not be only one such place or portal. With the example of Symbiota portals, we note that there can be many such places, each used by a particular research community, but none necessarily regulating the others, especially in the governance of metadata standards. Second is provisioning adequate infrastructure to manage and provide access to portals, datasets, and metadata standards. Depending on the research project, the relevant infrastructure may be a portal that is much smaller than global portals, or one that has the capacity to enable users to iteratively update data records based on previous work of cleaning the relevant data. Third is governing the resulting resource for the benefit of a community of users and stakeholders beyond a single research project or lab.

We further propose two criteria by which to evaluate digital datasets in portals. These criteria can be used to evaluate portals for their capacities to produce pooled data, and especially for determining if the pool provides widespread benefit to a community of users and stakeholders. The first criterion is for multiscalability, which holds that pooled datasets retain accurate metadata labels across fine to coarse-grained partitions within any given metadata category. This criterion means that when datasets are pooled, they'll carry with them few false negative and false positive annotations. In biodiversity, partitions in a metadata category can include taxonomic names in different levels of linnaean hierarchy, and wider or smaller geographic ranges. The second criterion is for tunability, which holds that it's possible to generate pooled datasets from portals that are fit for use to answer a range of types of problems. This criterion means that, for the governance of metadata standards for a portal, one research agenda doesn't swamp out other research agendas held within the community of users. Global portals aim to characterize trends in species abundance, population sizes, and geographic ranges at intercontinental and global scales. When researchers with more focused aims want to use the latest data from such portals, they must download the latest data and spend significant time cleaning it, work that is ultimately lost to other potential users and must be repeated without reward for each new data release. This work is lost because the global portals lack the capacities to incorporate it. We presented a case study of the Symbiota platform for biodiversity data portals as an exemplar for how data pooling can support multiscale, tunable datasets without necessitating consensus or strong unification of classificatory theories, problems, or methodologies across fields.

More generally, practices of data pooling represent a novel context for uncovering principles that characterize in which situations and how integration *without* unification can offer a *better* regulative ideal for scientific practice than integration with unification. While studies of explanatory integration have focused on complexity and local contingencies as obstacles to

unification, philosophers have largely overlooked connections with practices of producing integrated datasets. Brigandt (2010), for example, argues that explanatory integration is driven by solving specific problems, but data pooling in biodiversity research is motivated by the desire to address a wide range of problems and scientific aims. In particular, we hypothesize that the existence of general problem types with standardized "normal forms" of solutions (Culp and Kitcher 1989) is a major factor driving infrastructure development and data pooling efforts. Of special interest would be cases where these problem types reflect competing social or policy perspectives.

One example is the debate over whether conservation policy is most effective if we classify species into larger units under the biological species concept or smaller into units under the phylogenetic concept (Zachos 2018). Under the premise that data unification is required for global conservation of species, consensus on the choice of species concepts is necessary. We've shown, however, that this premise is sometimes false, and that by abandoning it when appropriate, researchers can dissolve disagreements about the "best" ways of characterizing and prioritizing phenomena for different purposes and constituent groups.

References

Araújo, Miguel, Robert Anderson, et al. 2019. "Standards for Distribution Models in Biodiversity Assessments." *Science Advances* 5: eaat4858.

Aronova, Elena, Karen Baker, and Naomi Oreskes. 2010. "Big Science and Big Data in Biology" *Historical Studies in the Natural Sciences* 40: 183–224.

Baker, Karen, and Florence Millerand. 2016. "Infrastructuring Ecology: Challenges in Achieving Data Sharing." In *Collaboration in the New Life Sciences*, 133–60. Routledge.

Berman, F, R Rutenbar, et al. 2016. "Realizing the Potential of Data Science." NSF Computer and Information Science and Engineering Advisory Committee.

Bisby, Frank A. 2000. "The Quiet Revolution: Biodiversity Informatics and the Internet." *Science* 289 (5488): 2309–12.

Bowker, Geoffrey. 2000. "Biodiversity Datadiversity." *Social Studies of Science* 30: 643–83.

Brigandt, Ingo. 2010. "Beyond Reduction and Pluralism." *Erkenntnis* 73: 295–311.

Campbell, Dana L., Anne E. Thessen, and Leslie Ries. 2020. "A Novel Curation System to Facilitate Data Integration across Regional Citizen Science Survey Programs." *PeerJ* 8 (July): e9219.

Culp, Sylvia, and Philip Kitcher. 1989. "Theory Structure and Theory Change in Contemporary Molecular Biology." *British Journal for the Philosophy of Science* 40: 459–83.

de Jong, Yde, Juliana Kouwenberg, Louis Boumans, et al. 2015. "PESI - a Taxonomic Backbone for Europe." *Biodiversity Data Journal* 3: e5848.

Devictor, Vincent, and Bernadette Bensaude-Vincent. 2016. "From Ecological Records to Big Data." *History and Philosophy of the Life Sciences* 38.

Enquist, Brian, Xiao Feng, et al. 2019. "The Commonness of Rarity: Global and Future Distribution of Rarity Across Land Plants." *Science Advances* 5: eaaz0414.

Franz, Nico, and Beckett Sterner. 2018. "To Increase Trust, Change the Social Design Behind Aggregated Biodiversity Data." *Database*. doi:10.1093/database/bax100.

Gadelha Jr., Luiz M. R., et al. 2018. "A Survey of E-Biodiversity: Concepts, Practices, and Challenges." ArXiv:1810.00224 [Cs, q-Bio], September. http://arxiv.org/abs/1810.00224.

Garnett, Stephen T., Les Christidis, Stijn Conix, Mark J. Costello, Frank E. Zachos, Olaf S. Bánki, Yiming Bao, et al. 2020. "Principles for Creating a Single Authoritative List of the World's Species." *PLOS Biology* 18 (7): e3000736.

Gerson, Elihu M. 1983. "Scientific Work and Social Worlds." *Science Communication* 4: 357–77.

Gesing, Sandra, Nancy Wilkins-Diehr, et al. 2017. "Science Gateways: The Long Road to the Birth of an Institute." Hawaii International Conference on System Sciences. doi:10.24251/HICSS.2017.755.

Global Biodiversity Information Facility. https://www.gbif.org/. Accessed July 28, 2020.

Godfray, HCJ. 2002. "Challenges for Taxonomy." *Nature* 417 (6884): 17–19.

Grantham, Todd A. 2004. “Conceptualizing the (Dis)Unity of Science.” *Philosophy of Science* 7: 133–55.

Gries, Corinna, Edward Gilbert, and Nico Franz. 2014. “Symbiota – a Virtual Platform for Creating Voucher-Based Biodiversity Information Communities.” *Biodiversity Data Journal*. doi:10.3897/BDJ.2.e1114.

Guala, Gerald F. 2016. “The Importance of Species Name Synonyms in Literature Searches.” *PLOS ONE* 11 (9): e0162648.

Guralnick, Robert, Andrew Hill, and Meredith Lane. 2007. “Towards a Collaborative, Global Infrastructure for Biodiversity Assessment.” *Ecology Letters* 10: 663–72.

Guralnick, Robert, Ramona Walls, and Walter Jetz. 2018. “Humboldt Core.” *Ecography* 41: 713–25.

Han, Xuemei, Carmen Josse, et al. 2017. “Monitoring National Conservation Progress with Indicators Derived From Global and National Datasets.” *Biological Conservation* 213: 325–34.

Hobern, Donald, Brigitte Baptiste, et al. 2019. “Connecting Data and Expertise: A New Alliance for Biodiversity Knowledge.” *Biodiversity Data Journal* 7: e33679.

Leonelli, Sabina. 2013. “Integrating Data to Acquire New Knowledge: Three Modes of Integration in Plant Science.” *Studies in the History and Philosophy of Biological and Biomedical Sciences* 44: 503–14.

Leonelli, Sabina. 2016. *Data-Centric Biology*. Chicago: University of Chicago Press.

Leonelli, Sabina. 2019. “The Challenges of Big Data Biology.” *eLife* 8: 69.

Mesibov, Robert. 2013. “A Specialist’s Audit of Aggregated Occurrence Records.” *ZooKeys* 293: 1–18.

Mesibov, Robert. 2018. “An Audit of Some Processing Effects in Aggregated Occurrence Records.” *ZooKeys* 751: 129–46.

Mitchell, Sandra. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

O’Malley, Maureen, and Orkun Soyer. 2012. “The Roles of Integration in Molecular Systems Biology.” *Studies in the History and Philosophy of Biological and Biomedical Sciences* 43: 58–68.

Pecl, Gretta T., et al. 2017. “Biodiversity Redistribution under Climate Change: Impacts on Ecosystems and Human Well-Being.” *Science* 355 (6332): eaai9214. Remsen, David. 2016. “The Use and Limits of Scientific Names in Biological Informatics.” *ZooKeys* 550: 207–23.

Peterson, A Townsend, Sandra Knapp, Robert Guralnick, et al. 2010. “The Big Questions for Biodiversity Informatics.” *Systematics and Biodiversity* 8: 159–68.

Ruggiero, Michael A, Dennis P Gordon, Thomas M Orrell, et al. 2015. “A Higher Level Classification of All Living Organisms.” *PLoS ONE* 10: e0119248.

Star, Susan Leigh, and James Griesemer. 1989. “Institutional Ecology, ‘Translations’ and Boundary Objects.” *Social Studies of Science* 19: 387–420.

Sterner, Beckett, and Nico Franz. 2017. “Taxonomy for Humans or Computers?” *Biological Theory* 12: 99–111.

Sterner, Beckett, Joeri Witteveen, and Nico Franz. 2020. “Coordination Instead of Consensus Classifications.” *History and Philosophy of the Life Sciences*. doi:10.1007/s40656-020-0300-z

Turnhout, Esther, and Susan Boonman-Berson. 2011. “Databases, Scaling Practices, and the Globalization of Biodiversity.” *Ecology and Society* 16.

Urban, M. C., et al. 2016. “Improving the Forecast for Biodiversity under Climate Change.” *Science* 353: aad8466–aad8466.

Wieczorek, John, David Bloom, et al. 2012. “Darwin Core.” *PLoS ONE* 7: e29715.

Zachos, Frank E. 2018. “Mammals and Meaningful Taxonomic Units: The Debate about Species Concepts and Conservation.” *Mammal Review* 48: 153–59.